Exome sequencing identifies a novel multiple sclerosis susceptibility variant in the *TYK2* gene

David A. Dymant, DPhil,
MD
M. Zameel Cader, MD,
DPhil
Michael J. Chao, DPhil
Matthew R. Lincoln,
DPhil, MD
Katie M. Morrison, MSc
Giulio Disanto, MD
Julia M. Morahan, PhD
Gabriele C. De Luca,
MD, DPhil
A. Dessa Sadovnick, PhD
Pierre Lepage, PhD
Alexandre Montpetit, PhD
George C. Ebers, MD,
FMedSci
Sreeram V. Ramagopalan,
DPhil

Correspondence & reprint
requests to Dr. Ebers:
george.ebers@clneuro.ox.ac.uk

ABSTRACT

Objective: To identify rare variants contributing to multiple sclerosis (MS) susceptibility in a family we have previously reported with up to 15 individuals affected across 4 generations.

Methods: We performed exome sequencing in a subset of affected individuals to identify novel variants contributing to MS risk within this unique family. The candidate variant was genotyped in a validation cohort of 2,104 MS trio families.

Results: Four family members with MS were sequenced and 21,583 variants were found to be shared among these individuals. Refining the variants to those with 1) a predicted loss of function and 2) present within regions of modest haplotype sharing identified 1 novel mutation (rs55762744) in the tyrosine kinase 2 (*TYK2*) gene. A different polymorphism within this gene has been shown to be protective in genome-wide association studies. In contrast, the *TYK2* variant identified here is a novel, missense mutation and was found to be present in 10/14 (72%) cases and 28/60 (47%) of the unaffected family members. Genotyping additional 2,104 trio families showed the variant to be transmitted preferentially from heterozygous parents (transmitted 16: not transmitted 5; $\chi^2 = 5.76$, $p = 0.016$).

Conclusions: Rs55762744 is a rare variant of modest effect on MS risk affecting a subset of patients (0.8%). Within this pedigree, rs55762744 is common and appears to be a modifier of modest risk effect. Exome sequencing is a quick and cost-effective method and we show here the utility of sequencing a few cases from a single, unique family to identify a novel variant. The sequencing of additional family members or other families may help identify other variants important in MS. *Neurology*® 2012;79:406-411

GLOSSARY

bp = base pair; GERP = genomic evolutionary rate profiling; MS = multiple sclerosis; SNP = single nucleotide polymorphism.

Multiple sclerosis (MS) is a common, complex, neurologic condition with both environmental and genetic factors contributing to risk. Epidemiologic studies have highlighted the significant influence of latitude on MS prevalence and Epstein-Barr virus infection, smoking, and low vitamin D levels are strongly implicated environmental risk factors.^{1,2} Genetic contributions are clear from family studies with 15%–20% of individuals with MS having an affected relative.³ Furthermore, monozygotic twins have a 25% concordance rate as compared to the 5% seen in dizygotic twins.⁴ These and other familial studies have demonstrated that genes contribute significantly to the familial aggregation of MS.

Despite a strong genetic component, it is extremely rare to find families with 5 or more individuals with MS in successive generations. We previously reported a family with up to 15 individuals with MS.⁵ MS was present in 4 generations and penetrance was relatively consistent

Editorial, page 396

From The Department of Medical Genetics (D.A.D.), University of Ottawa, Ottawa, Canada; Medical Research Council Functional Genomics Unit, Department of Physiology, Anatomy and Genetics (M.Z.C.), Nuffield Department of Clinical Neurosciences (Clinical Neurology) (M.Z.C., M.J.C., K.M.M., G.D., J.M.M., G.C.D., G.C.E., S.V.R.), and Wellcome Trust Centre for Human Genetics (M.J.C., M.R.L., K.M.M., G.D., J.M.M., G.C.D., G.C.E., S.V.R.), University of Oxford, Oxford, UK; Department of Medical Genetics (A.D.S.) and Faculty of Medicine, Division of Neurology (A.D.S.), University of British Columbia, Vancouver; and McGill University and Genome Québec Innovation Centre (P.L., A.M.), Montreal, Quebec, Canada.

Study funding: This work was funded by the Multiple Sclerosis Society of Great Britain and Northern Ireland, the Multiple Sclerosis Society of Canada Scientific Research Foundation, and the Wellcome Trust (grant 075491/Z/04).

Go to Neurology.org for full disclosures. Disclosures deemed relevant by the authors, if any, are provided at the end of this article.

between these generations. The family appeared to implicate a monogenic form of MS but a genome-wide screen for linkage showed no significant linkage.⁶ There were a number of regions that showed mild, nonsignificant, evidence for linkage in this family. We speculated that the presence of phenocopies or genetic heterogeneity may hinder the ability to identify a significantly linked locus in this unique family.⁶

Massively parallel (next-generation) sequencing has been responsible for the identification of several autosomal recessive and dominant genes in Mendelian disease.⁷ Given the relatively high penetrance and apparent autosomal dominant segregation pattern observed in the pedigree, we applied next-generation sequencing technology to 4 individuals with MS in the pedigree.

METHODS **Standard protocol approvals, registrations, and patient consents.** The institutional review board of the University of Western Ontario approved this study.

Patient ascertainment. Ascertainment and a detailed clinical description of the family has been described in detail elsewhere.^{5,6} There is a wide range of age at disease onset with the average onset age being 27 years (17–49 years). The female to male ratio is 1.8:1 and all forms of MS are represented in the pedigree (i.e., relapsing-remitting, primary progressive, and secondary progressive). The family is of German ancestry and there is no consanguinity. The pedigree is shown in figure 1.

Exome sequencing and analysis. DNA was extracted by standard methods. Four individuals were randomly selected for exome sequencing and are shown in the pedigree (figure 1). Exome capture was performed with the SureSelect Human All Exon kit (Agilent Technologies). Exon-enriched DNA was sequenced by the Illumina Genome Analyzer II platform following the manufacturer's instructions (Illumina). Raw image files were processed by the Illumina pipeline (version 1.3.4) for base calling with default parameters. The sequencing reads were aligned to the NCBI human reference genome (NCBI36.3) using bowtie 0.12.7⁸ (options “-a –best –strata”). Single nucleotide polymorphisms (SNPs) were subsequently called using SAMTools

(v0.1.8 [r613]⁹) with the following quality criteria: $\geq \times 20$ coverage and a Phred-like¹⁰ consensus quality of ≥ 30 and a SNP quality score of ≥ 100 . Variants were defined as heterozygous when $\geq 25\%$ of all nucleotides at the position showed nonreference bases and as homozygous according to the initial SAMTools classification (using default parameters). We used PolyPhen2 to assess nonsynonymous variants for a likely functional impact¹¹; both the HumDiv and HumVar datasets were screened.

Reads for which bowtie was unsuccessful in identifying an ungapped alignment were mapped to the human reference genome with BWA,¹² using default parameters. Indels of up to 4 bases were subsequently called using SAMTools (v0.1.8 [r613]⁹) with the following quality criteria: $\geq \times 20$ coverage, a Phred-like¹⁰ consensus quality of ≥ 50 , and a SNP quality score of ≥ 100 . Variants were defined as heterozygous when $\geq 25\%$ of all nucleotides at the position showed nonreference bases and as homozygous according to the initial SAMTools classification (using default parameters).

Linkage analysis and validation SNP genotyping. Files for linkage analyses were prepared with MEGA2¹³ and parametric linkage was performed with SIMWALK 2.0.¹⁴ A disease allele frequency was set at 0.001, the phenocopy rate was 0.001, and the penetrance was set low (0.01) to permit a more inclusive, albeit less powerful, screen for linkage. The analyses tested linkage to 681 genotyped markers with mapped distances taken from the Marshfield Genotyping Centre Database described in detail elsewhere.⁶ The analyses also included genotypes from the *HLA-DRB1* locus presented in the original publication.⁵

An additional cohort of 2,104 parents–affected child trios and 1,543 healthy controls were available for confirmation studies. Genotyping of additional families was performed using TaqMan (Applied Biosystems) assays.

RESULTS Paired, 76 base pair (bp) reads from postenrichment shotgun libraries were aligned to the reference genome. On average, 3.1 gigabases (Gb) of mapped sequence was generated per individual and 85% of reads were mapped. The average coverage of each exome was 70-fold. On average, 57,750 SNPs were called per individual (range 53,417–62,098), of which 64% were already annotated in a public database (dbSNP v131). A total of 19,743 SNPs were in common among all 4 patients and 7,701 of these were novel. A total of 396 SNPs were predicted to be damaging (loss of function) by PolyPhen2, of which

Figure 1 Pedigree in family with high prevalence of multiple sclerosis

Figure 2 Filtering of variants

240 had a dbSNP ID. On average, 5,038 indels were called per individual, of which 13.5% were previously annotated in dbSNP v131. A total of 1,840 indels were in common among all 4 patients, 1,562 were novel (i.e., not present in dbSNP), and 367 were exonic. In order to refine our list of variants, we included only those variants that were frameshift indels or PolyPhen2 predicted to be loss of function (damaging) and were common to all 4 affected individuals and present in regions that showed evidence for linkage under a “rare variant” model of multi-point linkage analyses (figure 2; table).

A single SNP in the *TYK2* gene on chromosome 19p13 was the only variant identified using these filter conditions (figure 2; table). The variant encodes a missense mutation in exon 3 of *TYK2* that changes an alanine to threonine (A53T) and it is predicted to

impact protein function by PolyPhen2 and SIFT.¹⁵ The genomic evolutionary rate profiling (GERP) score was +4.20. The GERP scores for rs897738 and rs6427384, the 2 “possibly” damaging variants, were +2.68 and −1.62, respectively (table). The A53T residue is highly conserved between species.

The variant at *TYK2* has previously been reported once in dbSNP by a study on cancer genomes and has an ID of rs55762744. The SNP was not reported by the 1000 Genomes project. MS has previously been shown to be associated with another, nonsynonymous SNP (rs34536443) in *TYK2* (P1104A), which lies over 25 kb downstream of rs55762744. The 4 individuals sequenced did not carry the rs34536443 MS-associated SNP. There were no obvious differences with MS phenotype or HLA genotype in the *TYK2* A53T variant carriers and noncarriers.

The variant was genotyped in the remaining affected family members. Of the affected individuals with DNA available, a total of 10/14 (72%) were positive for the *TYK2* mutation. The unaffected family members were also genotyped and a total of 28/60 (47%) were positive for the variant. When a parent was a heterozygous carrier for the mutation, the parent transmitted to an affected offspring 10 times and did not transmit 2 times ($\chi^2 = 5.33$, $p = 0.02$). We then genotyped 2,104 MS “trio” families and 16 cases were positive for the *TYK2* variant (0.8%). Twenty-one unaffected parents were heterozygous for rs55762744 and the risk allele was transmitted preferentially to 16 cases and not transmitted to 5 cases ($\chi^2 = 5.76$, $p = 0.016$). A control sample was genotyped for the variant and 10 individuals of 1,543 (0.6%) were positive for the T allele and not significantly different from the affected allele frequency.

We also tested for any rare variants in the other known MS susceptibility genes outside of the linkage regions identified by the MS GWA studies. These genes included *IL2RA*, *IL7R*, *CLEC16A*, *CD58*, *TNFRSF1A*, *IRF8*, *EV15*, *KIF1B*, *KIF21B*, *CD40*, *STAT3*, *CBLB*, *CD6*, *CD226*, and *GPC5*. There were no frameshift insertions, deletions, or damaging SNPs identified in any of the 4 sequenced individuals at these genes.

DISCUSSION The family presented here is unique for its number of affected individuals with MS in multiple generations, the extent of clinical investigation performed, and the advantage of 20 years of routine follow-up and prospectively identified cases. The family appeared to be segregating MS as an autosomal dominant condition and this is suggestive a priori that this family may harbor a dom-

Table Number of “probably” damaging variants identified in regions of linkage

Linkage region ^a	LOD ≥ 1	No. of genes	No. of shared and damaging variants ^a
1p13-q23	1.09	690	0
2q13-q21	1.08	238	0
7q36-tel	1.28	173	0
16q23-q24	0.99	89	0
19p13	1.75	334	1

^a Missense/nonsense and frameshift indels. An additional 2 variants, not shown here, were called as “possibly damaging” by PolyPhen2. One non-novel variant, rs6427384, is located at 1p13-q23 and the second “possibly damaging” variant is rs897738.

inant acting mutation and be amenable to a resequencing strategy.

The sequencing of 4 individuals from a single family resulted in the identification of a great number of shared variants (>20,000). The selection of cases is an important step in family studies of this type. Given the absence of significant linkage (lod > 3.3) we opted to select the individuals at random. If linkage had been observed, the individuals selected for sequencing would have been best selected as those sharing a common haplotype. Alternatively, sequencing family members that are relatively more distantly related may be overly conservative given the likely genetic heterogeneity present in a family with no significant linkage. The filtering of the variants can be a delicate process; the challenge is to reduce the number of variants to a manageable number of candidates and yet not to lose a possible causative variant in the process. Including the linkage results under a “rare” disease allele model was integral to the filtering process and has been used with success in Mendelian disease.¹⁶ We did not observe significant linkage and we assessed regions with mild to moderate evidence of haplotype sharing (i.e., lod ≥ 1). We were confident of the result for many reasons; *TYK2* is a gene that has recently been shown to be implicated in MS as a protective allele (odds ratio = 0.8) in several large association studies.^{17–19} Furthermore, the observation of a GERP score of +4.20 was also in keeping with a deleterious genetic change.²⁰ Polyphen2 and SIFT also showed evidence of an impact on protein function and, finally, there was significant transmission disequilibrium from parents to affected off-

spring (16 times transmitted vs 5 times not transmitted). A large, family-based TDT cohort for follow-up was necessary as the T allele was relatively common in our control sample (0.6%) and a case-control approach would not have easily identified this allele as a susceptibility variant. There may be other rare variants at *TYK2* that modify risk and resequencing of the entire gene in hundreds to thousands of cases may prove useful in their identification. It is very likely that our conservative filtering approach missed other non-*TYK2* risk alleles segregating within this family. The resequencing of the exome of other or all family members may be of benefit; in particular, those lacking the *TYK2* risk variant.

TYK2 encodes a protein kinase that operates to phosphorylate proteins in the JAK-STAT3 immune pathway and regulate downstream immune cytokines.²¹ Analysis of protein–protein interactions by STRING²² highlighted an interaction of *TYK2* with many of the associated MS genes identified by GWA studies (figure 3) including *IL7R* and *STAT3*, lending further support for the involvement of *TYK2* in MS risk and likely acting through an immune mechanism. It is intriguing that the previously known MS-associated SNP, rs34536443, is considered a protective allele, whereas we provide evidence that rs55762744 confers risk to MS within this family. A recent study has shown the protective allele to alter tyrosine kinase activity and modify the cytokine profile to a T_H2 phenotype.²¹ It may be common to see both protective and susceptibility alleles at the GWAS-associated genes akin to the variants observed

Figure 3 Multiple sclerosis susceptibility protein–protein interactions

(A) Analysis performed using STRING.²⁴ Thicker lines indicate more confident protein–protein interactions. (B) The protein structure of *TYK2* with the familial variant (Ala53) identified by exome sequencing.

at the MHC. An example would be *HLA-DRB1*15* that acts as a susceptibility allele while *HLA-DRB1*14* acts as a protective allele.²³ How the newly identified variant acts to increase MS risk is not known and functional studies are necessary and currently underway to clarify its role.

The most appropriate method for the identification of rare variants in complex, qualitative disease is being debated.²⁴ A suggested approach involves the exome or genome sequencing of many affected individuals and healthy controls. Having enough power and resources may be problematic at this time. Alternatively, the enrichment and resequencing of select regions that have been implicated by GWA studies may be fruitful. This latter strategy would have worked for the present study with (or without) the modest linkage information included as a filter. From this report, it is clear that a rare variant can have low penetrance even in a seemingly autosomal dominant family and this can confound linkage analysis and subsequent filtering if unaffected cases are included. In our pedigree 75% of carriers of the *TYK2* mutation did not have MS and the linkage was not significant (lod = 1.75) although 19p13 was the region with the strongest evidence of haplotype sharing in a scan of the entire genome. While linkage worked for this investigation we also benefited from previous GWA studies that identified *TYK2* as the site of a resistance allele.

We report here one of the first exome sequencing strategies to identify a novel variant for a common complex disease and we believe this approach will be of significant utility for many other traits in the near future.

AUTHOR CONTRIBUTIONS

D. Dymant: drafting/revising the manuscript, study concept or design, analysis or interpretation of data, statistical analysis, study supervision, obtaining funding. M.Z. Cader: drafting/revising the manuscript, study concept or design, analysis or interpretation of data. M. Chao: drafting/revising the manuscript, acquisition of data. M. Lincoln: analysis or interpretation of data, acquisition of data. K. Morrison: drafting/revising the manuscript, contribution of vital reagents/tools/patients, study supervision. G. Disanto: analysis or interpretation of data. J. Morahan: drafting/revising the manuscript. G. DeLuca: drafting/revising the manuscript. A.D. Sadovnick: drafting/revising the manuscript, study concept or design, acquisition of data, study supervision, obtaining funding. P. Lepage: analysis or interpretation of data, acquisition of data. A. Montpetit: drafting/revising the manuscript, acquisition of data, study supervision. G. Ebers: drafting/revising the manuscript, study concept or design, analysis or interpretation of data, contribution of vital reagents/tools/patients, study supervision. S. Ramagopalan: drafting/revising the manuscript, study concept or design, analysis or interpretation of data, acquisition of data, statistical analysis, study supervision.

DISCLOSURE

D. Dymant receives support from the CIHR Institute of Genetics Clinical Investigatorship Award. M.Z. Cader receives research support from the Medical Research Council of the United Kingdom. M. Chao, M. Lincoln, K. Morrison, G. Disanto, J. Morahan, and G. De Luca report no disclo-

tures. A.D. Sadovnick has received funding for travel from Bayer Canada, the European Charcot Foundation, the MS Society of Canada, the National MS Society, and the Consortium of MS Clinics; has received speaker honoraria from Teva Scientific and the MS Society of Canada; and has received research support from the MS Society of Canada Scientific Research Foundation (Co-PI) and the CIHR (Co-PI). P. Lepage and A. Montpetit report no disclosures. G. Ebers has received a speaker honorarium from Roche and a consultation fee from UCB Pharma and receives research support from Bayer Schering Pharma, the Multiple Sclerosis Society of the United Kingdom (grant 865_07 [PI] and grant 875_07 [PI]), and the Multiple Sclerosis Society of Canada Scientific Research Foundation (PI and co-PI). S. Ramagopalan receives research support from the Multiple Sclerosis Society of Canada Scientific Research Foundation and the Multiple Sclerosis Society of the United Kingdom. **Go to Neurology.org for full disclosures.**

Received July 14, 2011. Accepted in final form December 29, 2011.

REFERENCES

- Ascherio A, Munger KL. Environmental risk factors for multiple sclerosis: part I: the role of infection. *Ann Neurol* 2007;61:288–299.
- Ascherio A, Munger KL. Environmental risk factors for multiple sclerosis: part II: noninfectious factors. *Ann Neurol* 2007;61:504–513.
- Ebers GC, Koopman WJ, Hader W, et al. The natural history of multiple sclerosis: a geographically based study: 8: familial multiple sclerosis. *Brain* 2000;123:641–649.
- Willer CJ, Dymant DA, Risch NJ, Sadovnick AD, Ebers GC. Twin concordance and sibling recurrence rates in multiple sclerosis. *Proc Natl Acad Sci USA* 2003;100:12877–12882.
- Dymant DA, Cader MZ, Willer CJ, Risch N, Sadovnick AD, Ebers GC. A multigenerational family with multiple sclerosis. *Brain* 2002;125:1474–1482.
- Dymant DA, Cader MZ, Herrera BM, et al. A genome scan in a single pedigree with a high prevalence of multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2008;79:158–162.
- Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009;461:272–276.
- Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10:R25.
- Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
- Ewing B, Green P. Base-calling of automated sequencer traces using phred II: error probabilities. *Genome Res* 1998;8:186–194.
- Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248–249.
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
- Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill WP, Weeks DE. Mega2: data-handling for facilitating genetic linkage and association analyses. *Bioinformatics* 2005;21:2556–2557.
- Sobel E, Lange K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* 1996;58:1323–1337.

15. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4:1073–1081.
16. Sobreira NL, Cirulli ET, Avramopoulos D, et al. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet* 2010;6:e1000991.
17. Mero IL, Lorentzen AR, Ban M, et al. A rare variant of the TYK2 gene is confirmed to be associated with multiple sclerosis. *Eur J Hum Genet* 2010;18:502–504.
18. Ban M, Goris A, Lorentzen AR, et al. Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. *Eur J Hum Genet* 2009;17:1309–1313.
19. Johnson BA, Wang J, Taylor EM, et al. Multiple sclerosis susceptibility alleles in African Americans. *Genes Immun* 2010;11:343–350.
20. Cooper GM, Goode DL, Ng SB, et al. Single nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods* 2010;7:250–251.
21. Couturier N, Bucciarelli F, Nurdin RN, et al. Tyrosine kinase 2 variant influences T lymphocyte polarization and multiple sclerosis susceptibility. *Brain* 2011;134:693–703.
22. Szklarczyk D, Franceschini A, Kuhn M, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2011;39:D561–D568.
23. Ramagopalan SV, Morris AP, Dymant DA, et al. The inheritance of resistance alleles in multiple sclerosis. *PLOS Genet* 2007;9:1607–1613.
24. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010;11:415–425.

Neurology® Podcast

A randomized trial of stress management for the prevention of new brain lesions in MS

At www.neurology.org, click on the “Download Latest Issue” link or “Subscribe Now” to subscribe to the RSS Feed.

This week’s podcast features an interview by Dr. Pearce Korb with Dr. David Mohr about his paper on stress management for the prevention of new brain lesions in MS.

NEW CME Opportunity: Listen to this week’s Neurology® Podcast and earn 0.5 AMA PRA Category 1 CME Credits™ by answering the multiple-choice questions in the online Podcast quiz.

Scan this code with your smart phone to directly access the Podcast

Neurology® Online CME Program

Earn CME while reading *Neurology*. This program is available only to online *Neurology* subscribers. Simply read the articles marked CME, go to www.neurology.org, and click on CME. This will provide all of the information necessary to get started. The American Academy of Neurology (AAN) is accredited by the Accreditation Council for Continuing Medical Education (ACCME) to sponsor continuing medical education for physicians. *Neurology* is planned and produced in accordance with the ACCME Essentials. For more information, contact AAN Member Services at 800-879-1960.