**Web Materials**

MATERIALS AND METHODS

Discovery study population

*Genotype and Imputation information.* Saliva was collected using Oragene kits. Whole blood was collected, processed and extracted for DNA using GentraPuregene protocol. The genotyped Kaiser data set consisted of 1,660 white individuals (1,069 cases and 591 controls) with 594,345 autosomal SNPs as previously described.[1](#_ENREF_1) Several quality control measures were utilized to prepare the data for imputation. Using the dbSNP resource, all SNP names and physical location were updated to the February 2009 human reference sequence (GRCh37). Strand orientation was assessed using HapMap data for Utah residents with Northern and Western European ancestry (CEU) as extracted from the HapMart data bank (<http://hapmap.ncbi.nlm.nih.gov/biomart/martview/>). Alleles were flipped accordingly, when possible, to match the alignment of the top strand of the CEU data. Autosomal SNPs with minor allele frequencies (MAF) >0.45 in the control population were excluded due to concerns regarding ambiguous strand orientation. Furthermore, SNPs where the MAF difference between the control population and CEU was ≥20%, that were not in Hardy Weinberg Equilibrium (*P*<1x10-5), had missing genotypes >20%, or had MAF <0.0001 were excluded. Individuals with more than 10% missing genotypes were also excluded. This data set included 1,619 individuals (1,032 cases, 587 controls) with 490,633 SNPs. The data were imputed using the IMPUTE2 methodology with a pre-phasing step to produce estimated haplotypes using an accurate approach as implemented in ShapeIT, and using the 1000 Genomes Phase I integrated variant set (February, 2012) that included reference data (approximately 37 million SNPs) for 1,092 individuals from Africa, Asia, Europe, and the Americas.[2-5](#_ENREF_2) The pre-phased haplotypes were imputed in 5Mb chunks that did not span the centromere. Missing genotypes at typed SNPs were imputed if the maximum probability (calling threshold) was below 90%. After imputation, each SNP is represented by a set of three probabilities which correspond to the allele pairs (e.g. AA, AB, BB). Pedigree files were generated from the imputed data using GTOOL (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html), and a 90% probability at an individual genotype was required, otherwise the genotype was considered missing in the output files.

Lastly, SNPs that were imputed with <30% certainty (a metric similar to r2, which indicates that the amount of data at the imputed SNP is approximately equivalent to a set of perfectly observed genotype data in a sample size) were excluded from the final imputed data set. The imputed data set consisted of 17.5 million autosomal SNPs in 1,619 individuals. The presence of relatedness and population substructure was determined using multidimensional scaling as implemented in PLINKv1.07.[6](#_ENREF_6) We observed genetically related individuals (6 pairs of half-siblings and siblings; proportion of alleles shared across the genome >0.25), which were removed from all analyses. There were three dimensions that captured population structure in the dataset (C1, C2, and C3). Individuals with C1>0 were removed from genetic analyses, resulting in 1,016 MS cases and 583 controls with dense genetic data, of which 1,012 MS cases and 576 controls had non-missing exposure information. We subsequently restricted our analysis to imputed SNPs for which there was >90% threshold calling for a single genotype, and >90% imputation certainty.

Statistical Power

We investigated the power to detect GxE interactions within the discovery (Kaiser) data. Odds ratios of the interaction (ORIs) ranging from 0.1 to 3.0 were examined, baseline risk of 0.1%, marginal genetic effect of 1, environmental exposure of 40% (Table 1B), and a marginal environmental effect of 1.35 (Table 1B). We assumed a two-sided type 1 error of 5% (α=0.05), as this study was not hypothesis free, but strongly hypothesis driven, focusing on a small number of candidate genes, and one environmental exposure. Results indicated that our investigation had sufficient power (>80%) to detect GxE interaction ORs ˂0.5 and ˃1.9 for variants with minor allele frequencies >10%, GxE interaction ORs <0.6 and >1.6 for variants with minor allele frequencies >20% and GxE interaction ORs <0.7 and >1.5 for variants with minor allele frequencies >30% (data not shown). Therefore, we restricted all analyses to variants with minor allele frequencies >10%.

Web Table 1A: Gene-environment interactions for common *NAT1* and *NAT2* phenotypes1

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Allele | Acetylation Status | OR-IXN | L95 | U95 | P | Cases | Controls | Haplotype | SNP1 | SNP2 | SNP3 |
| *NAT1\*3* | Neutral | 0.81 | 0.58 | 1.14 | 0.2206 | 0.2629 | 0.2664 | A | rs15561 |  |  |
| *NAT2\*5* | Slow | 0.98 | 0.73 | 1.31 | 0.8786 | 0.4311 | 0.4417 | C | rs1801280 |  |  |
| *NAT2\*5A* | Slow | 0.95 | 0.71 | 1.27 | 0.7317 | 0.3829 | 0.3945 | CT | rs1801280 | rs1799929 |  |
| *NAT2\*5B* | Slow | 0.94 | 0.70 | 1.26 | 0.6735 | 0.3691 | 0.3688 | CTG | rs1801280 | rs1799929 | rs1208 |
| *NAT2\*5C* | Slow | 0.97 | 0.72 | 1.30 | 0.8238 | 0.4173 | 0.416 | CG | rs1801280 | rs1208 |  |
| *NAT2\*5D* | Slow | 0.98 | 0.73 | 1.31 | 0.8786 | 0.4311 | 0.4417 | A | rs1801280 |  |  |
| *NAT2\*6* | Slow | 1.15 | 0.83 | 1.60 | 0.3995 | 0.2968 | 0.2933 | A | rs1799930 |  |  |
| *NAT2\*6A* | Slow | 1.20 | 0.86 | 1.66 | 0.2849 | 0.2781 | 0.277 | TA | rs1041983 | rs1799930 |  |
| *NAT2\*6B* | Slow | 1.15 | 0.83 | 1.60 | 0.3995 | 0.2968 | 0.2933 | A | rs1799930 |  |  |
| *NAT2\*11A* | Fast | 0.94 | 0.70 | 1.26 | 0.6721 | 0.4193 | 0.4245 | T | rs1799929 |  |  |
| *NAT2\*12A* | Fast | 0.94 | 0.70 | 1.27 | 0.7014 | 0.4144 | 0.4151 | G | rs1208 |  |  |
| *NAT2\*12C* | Fast | 0.93 | 0.70 | 1.25 | 0.6536 | 0.3696 | 0.3688 | TGA | rs1799929 | rs1208 |  |
| *NAT2\*13A* | Fast | 1.13 | 0.82 | 1.55 | 0.4521 | 0.3243 | 0.3242 | T | rs1041983 |  |  |

1 Logistic regression models adjusted for gender, year of birth, population ancestry, and college education.

Web Table 1B: Gene-environment interactions by acetylation phenotypes 1

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Acetylation status | Collapsed Haplotypes | Haplotype | SNP1 | SNP2 | SNP3 | SNP4 | SNP5 | OR-IXN | L95 | U95 | P |
| *NAT1 fast* | *NAT1\*11* | ATTAG | rs15561 | rs4986988 | rs4986989 | rs4986990 | rs4986783 | 2.04 | 0.89 | 4.68 | 0.094 |
|  | *NAT1\*30* | A | rs4987076 |  |  |  |  |  |  |  |  |
| *NAT1 slow* | *NAT1\*14* | AA | rs15561 | rs4986782 |  |  |  | 0.78 | 0.26 | 2.30 | 0.655 |
|  | *NAT1\*15* | T | rs5030839 |  |  |  |  |  |  |  |  |
|  | *NAT1\*17* | T | rs56379106 |  |  |  |  |  |  |  |  |
| *NAT2 slow* 2 | *NAT2\*5* | C | rs1801280 |  |  |  |  | 1.13 | 0.80 | 1.60 | 0.479 |
|  | *NAT2\*6* | A | rs1799930 |  |  |  |  |  |  |  |  |
|  | *NAT2\*7* | A | rs1799931 |  |  |  |  |  |  |  |  |
|  | *NAT2\*14* | A | rs1801279 |  |  |  |  |  |  |  |  |

1 Logistic regression models adjusted for gender, year of birth, population ancestry, and college education.

2 The NAT2 acetylator phenotype was determined by four SNPs: rs1801279 (191G>A), rs1801280 (341T>C), rs1799930 (590G>A) and rs1799931 (857G>A). Samples homozygous for the major allele (common) for all SNPs were classified as rapid acetylator phenotype, samples heterozygous for any one of the SNPs were classified as intermediate acetylator phenotype and samples homozygous variant for one or more SNPs or heterozygous for two or more SNPs were classified as slow acetylator phenotype.

Web Table 2: Marginal Association Between Tobacco Smoke Exposure and MS in a Matched Analysis

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Characteristic | Kaiser  Cases | Kaiser Controls | ORcrude  (95% CI) | *P* | ORadjusted  (95% CI) | *P* |
| Ever Smoker(%)a | 49.1 | 40.4 | 1.48  (1.14, 1.92) | **0.0033** | 1.40  (1.07, 1.82) | **0.014** |
| Smoker at age 20 (%) b | 50.2 | 41.9 | 1.92  (1.41, 2.63) | **3.9 x 10-5** | 1.88  (1.39, 2.59) | **9.9 x 10-5** |
| Smoker in the  reference year (%) c | 39.9 | 36.2 | 2.06  (1.37, 3.10) | **5.5 x 10-4** | 1.97  (1.29, 3.00) | **0.0016** |

Abbreviations: OR, odds ratio; CI, confidence interval

a Adjusted for college education and population ancestry

b N=521 cases, 522 controls

c N=462 cases, 463 controls

d N=416 cases, 417 controls

Web Table 3: Conditional Models a for Tobacco Smoke Exposure by the *NAT1* rs7388368 Genotype

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Study | rs7388368 GxE | | | | C/C | | | | A/C or C/C | | | |
|  | FA | FU | ORI  (95% CI) | *P* | FA | FU | OR  (95% CI) | *P* | FA | FU | OR  (95% CI) | *P* |
| Ever smoker b | 50.3 | 40.6 | 1.85  (1.15, 2.95) | **0.010** | 50.0 | 43.8 | 1.29  (0.81, 2.04) | 0.285 | 45.2 | 25.8 | 3.46  (1.28, 9.34) | **0.014** |
| Smoker at age 20 c | 34.9 | 22.6 | 1.43  (0.83, 2.45) | 0.195 | 31.0 | 25.8 | 1.23  (0.72, 2.12) | 0.452 | 30.0 | 15.0 | 3.24  (1.07, 9.80) | **0.037** |
| Smoker in the reference year d | 20.8 | 12.3 | 2.63  (1.23, 5.62) | **0.012** | 17.3 | 13.0 | 1.44  (0.67, 3.08) | 0.353 | 18.2 | 3.6 | 7.20  (1.39, 37.29) | **0.019** |

Abbreviation: ORI, odds ratio of interaction term; FA, frequency of smoking in cases (affected); FU, frequency of smoking in controls (unaffected); OR, odds ratio; CI, confidence interval

a Adjusted for college education and population ancestry

b N=487 matched pairs. In the stratified analyses there were: 178 matched pairs with the C/C genotype; 62 matched pairs with the A/C or A/A genotype.

c N=433 matched pairs where MS onset ≥20years in the case. In the stratified analyses there were: 155 matched pairs with the C/C genotype; 60 matched pairs with the A/C or A/A genotype.

|  |
| --- |
| d N=389 matched pairs. In the stratified analyses there were: 139 matched pairs with the C/C genotype; 55 matched pairs with the A/C or A/A genotype. |
|  |

Web Table 4:GxE Interaction Results for Ever/Never Smoker Status and *NAT1* Variants in Linkage Disequilibrium with rs7388368A (*r2*≥0.5) in at least one of the Study Populations.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SNP | Location | Allele | Kaiser study a | | | Swedish study b | | | Meta-analysisc | |
|  |  |  | *r2* | ORI (95% CI) | P | *r2* | ORI (95% CI) | P | ORI (95% CI) | *P* |
| rs1000214 | 18,053,598 | T | 0.50 | 1.80 (1.15, 2.81) | **0.010** | 0.40 | 1.34 (0.80, 2.24) | 0.27 | 1.52 (1.10, 2.11) | **0.011** |
| rs7388368 | 18,055,386 | A | 1 | 1.75 (1.19, 2.56) | **0.004** | 1 | 1.62 (1.05, 2.49) | **0.029** | 1.65 (1.25, 2.18) | **4.6 x 10-4** |
| rs4921877 | 18,055,584 | T | 0.85 | 1.46 (1.02, 2.09) | **0.039** | 0.86 | 1.39 (0.92, 1.54) | 0.12 | 1.40 (1.07, 1.83) | **0.013** |
| rs6586711 | 18,059,563 | C | 0.66 | 1.68 (1.21, 2.34) | **0.0022** | 0.74 | 1.22 (0.81, 1.83) | 0.34 | 1.42 (1.10, 1.83) | **0.0063** |
| rs4921578 | 18,062,043 | T | 0.59 | 1.26 (0.90, 1.76) | 0.17 |  |  |  |  |  |
| rs11997214 | 18,064,815 | G | 0.55 | 1.08 (0.78, 1.49) | 0.65 | 0.56 | 1.57 (1.07, 2.31) | **0.021** | 1.24 (0.97, 1.58) | 0.081 |
| rs7001893 | 18,065,337 | A | 0.65 | 1.19 (0.85, 1.68) | 0.31 |  |  |  |  |  |
| rs6586712 | 18,066,967 | G | 0.50 | 1.29 (0.93, 1.78) | 0.12 | 0.51 | 1.26 (0.87, 1.83) | 0.23 | 1.25 (0.98, 1.59) | 0.072 |
| rs7845127 | 18,067,395 | T | 0.64 | 1.26 (0.90, 1.76) | 0.18 | 0.63 | 1.57 (1.07, 2.31) | **0.022** | 1.37 (1.07, 1.75) | **0.014** |
| rs2410545 | 18,068,835 | T | 0.53 | 1.18 (0.85, 1.63) | 0.31 | 0.56 | 1.43 (0.98, 2.09) | 0.064 | 1.26 (0.99, 1.60) | 0.061 |

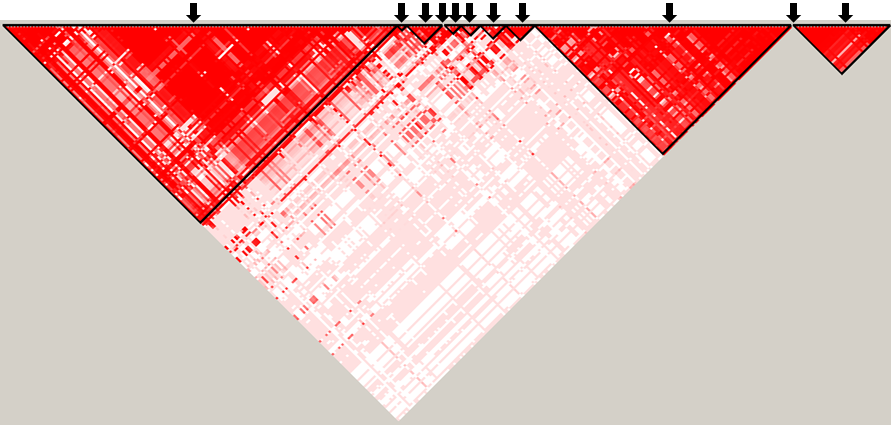
Abbreviation: ORI, odds ratio of interaction term; CI, confidence interval

a All models were adjusted for gender, year of birth, college education, and population ancestry.

b All models were adjusted for gender, year of birth, college education, and country of origin.

c Random-effects meta-analyses adjusted for gender, year of birth, and college education

Web Figure 1: Illustration of the extensive linkage disequilibrium (LD) across the 277 variants investigated, and the assignment of LD blocks based on a Solid Spine with a minimum D′ of 0.75. There are 11 independent regions, as identified by the black arrows.



Web Figure 2: Visualization of rs7388368 and the Broad ChromHMM track in UCSC genome browser, suggesting that it resides within an insulator.



REFERENCES

1. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE, Edkins S, Gray E, Booth DR, Potter SC, Goris A, Band G, Oturai AB, Strange A, Saarela J, Bellenguez C, Fontaine B, Gillman M, Hemmer B, Gwilliam R, Zipp F, Jayakumar A, Martin R, Leslie S, Hawkins S, Giannoulatou E, D'Alfonso S, Blackburn H, Martinelli Boneschi F, Liddle J, Harbo HF, Perez ML, Spurkland A, Waller MJ, Mycko MP, Ricketts M, Comabella M, Hammond N, Kockum I, McCann OT, Ban M, Whittaker P, Kemppinen A, Weston P, Hawkins C, Widaa S, Zajicek J, Dronov S, Robertson N, Bumpstead SJ, Barcellos LF, Ravindrarajah R, Abraham R, Alfredsson L, Ardlie K, Aubin C, Baker A, Baker K, Baranzini SE, Bergamaschi L, Bergamaschi R, Bernstein A, Berthele A, Boggild M, Bradfield JP, Brassat D, Broadley SA, Buck D, Butzkueven H, Capra R, Carroll WM, Cavalla P, Celius EG, Cepok S, Chiavacci R, Clerget-Darpoux F, Clysters K, Comi G, Cossburn M, Cournu-Rebeix I, Cox MB, Cozen W, Cree BA, Cross AH, Cusi D, Daly MJ, Davis E, de Bakker PI, Debouverie M, D'Hooghe M B, Dixon K, Dobosi R, Dubois B, Ellinghaus D, Elovaara I, Esposito F, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 2011;476(7359):214-9.

2. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics 2007;39(7):906-13.

3. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 2011;1(6):457-70.

4. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nature methods 2012;9(2):179-81.

5. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature genetics 2012;44(8):955-9.

6. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 2007;81(3):559-75.