frontiers in Genetics

ORIGINAL RESEARCH
published: 20 October 2016
doi: 10.3389/fgene.2016.00185

# Shared Genetic Etiology of Autoimmune Diseases in Patients from a Biorepository Linked to De-identified Electronic Health Records

*Nicole A. Restrepo*[1], *Mariusz Butkiewicz*[1], *Josephine A. McGrath*[2] *and Dana C. Crawford*[1,3]*

[1] *Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA,* [2] *Vanderbilt Eye Institute, Vanderbilt University Medical Center, Nashville, TN, USA,* [3] *Institute for Computational Biology, Case Western Reserve University, Cleveland, OH, USA*

Autoimmune diseases represent a significant medical burden affecting up to 5–8% of the U.S. population. While genetics is known to play a role, studies of common autoimmune diseases are complicated by phenotype heterogeneity, limited sample sizes, and a single disease approach. Here we performed a targeted genetic association study for cases of multiple sclerosis (MS), rheumatoid arthritis (RA), and Crohn's disease (CD) to assess which common genetic variants contribute individually and pleiotropically to disease risk. Joint modeling and pathway analysis combining the three phenotypes were performed to identify common underlying mechanisms of risk of autoimmune conditions. European American cases of MS, RA, and CD, ($n$ = 119, 53, and 129, respectively) and 1924 controls were identified using de-identified electronic health records (EHRs) through a combination of International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) billing codes, Current Procedural Terminology (CPT) codes, medication lists, and text matching. As expected, hallmark SNPs in MS, such as *DQA1* rs9271366 (OR = 1.91; $p$ = 0.008), replicated in the present study. Both MS and CD were associated with *TIMMDC1* rs2293370 (OR = 0.27, $p$ = 0.01; OR = 0.25, $p$ = 0.02; respectively). Additionally, *PDE2A* rs3781913 was significantly associated with both CD and RA (OR = 0.46, $p$ = 0.02; OR = 0.32, $p$ = 0.02; respectively). Joint modeling and pathway analysis identified variants within the KEGG NOD-like receptor signaling pathway and Shigellosis pathway as being correlated with the combined autoimmune phenotype. Our study replicated previously-reported genetic associations for MS and CD in a population derived from de-identified EHRs. We found evidence to support a shared genetic etiology between CD/MS and CD/RA outside of the major histocompatibility complex region and identified KEGG pathways indicative of a bacterial pathogenesis risk for autoimmunity in a joint model. Future work to elucidate this shared etiology will be key in the development of risk models as envisioned in the era of precision medicine.

**Keywords: autoimmune disease, Crohn's disease, rheumatoid arthritis, multiple sclerosis, electronic health records**

OPEN ACCESS

***Edited by:***
*Robert Klein,*
*Icahn School of Medicine at Mount Sinai, USA*

***Reviewed by:***
*Chunyan He,*
*Indiana University–Purdue University Indianapolis, USA*
*Xiaowei Sherry Yan,*
*Sutter Health, USA*

***Correspondence:**
*Dana C. Crawford*
*dana.crawford@case.edu*

***Specialty section:***
*This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics*

***Received:*** *14 June 2016*
***Accepted:*** *03 October 2016*
***Published:*** *20 October 2016*

***Citation:***
*Restrepo NA, Butkiewicz M, McGrath JA and Crawford DC (2016) Shared Genetic Etiology of Autoimmune Diseases in Patients from a Biorepository Linked to De-identified Electronic Health Records. Front. Genet. 7:185. doi: 10.3389/fgene.2016.00185*

## INTRODUCTION

Autoimmune (AI) and immune-mediated (IM) diseases are a driving force behind disability in the United States with 5–8% of the population affected (2002). Currently there are upwards of 80 conditions that occur as a consequence of immunological attacks on the body's tissues and organs. As a whole, AI and IM diseases share many epidemiological and pathogenic similarities. Most notable is the occurrence of a skewed sex distribution with females more likely to be affected by an AI disease. The female sex ratio of cases is most extreme in Sjögren syndrome (9:1) that occurs as a result of immune-mediated attack on salivary and lacrimal glands in comparison to type 1 diabetes which is sex neutral (Gale and Gillespie, 2001). Over the last decade, genome-wide association studies have consistently identified genes within the major histocompatibility complex (MHC), located on the short arm of chromosome 6, strong regulators of disease risk in a number of AI diseases such as multiple sclerosis (MS) (International Multiple Sclerosis Genetics Consortium et al., 2007, 2011; Patsopoulos et al., 2011), rheumatoid arthritis (RA) (Plenge et al., 2007; Raychaudhuri et al., 2008; Okada et al., 2012), and vitiligo (Jin et al., 2010; Quan et al., 2010). Despite these commonalties, their underlying genetic and molecular profiles vary and are hindered by low case counts with most AI diseases rarely occurring in the general population (i.e., Addison's disease prevalence estimated at ∼1 in 20,000). Elusive and overlapping symptoms further confound diagnosis of a condition, often leading patients on a long and costly diagnostic odyssey. Additionally, time to diagnosis, time between presentation of first symptoms to confirmed diagnosis, can approach a year in conditions such as amyotrophic lateral sclerosis (Paganoni et al., 2014).

Large-scale studies for AI and IM diseases typically require recruiting and consenting patients from specialty clinics and individual medical practices, all of which necessitates considerable financial investment and manpower. In recent years, researchers have begun to utilize electronic health records (EHR) for use in clinical and genetic association studies covering a range of conditions from type 2 diabetes (Ng et al., 2014) to response to drugs or treatment (Oetjens et al., 2014; Laper et al., 2016). Use of EHRs rapidly facilitates biomedical studies by providing researchers with access to a repository of extensive, longitudinal medical data. In particular, EHRs offer the opportunity to assess less common conditions such as MS by directly accessing data from specialty disease clinics within the framework of a health care organization. In conjunction with DNA repositories, it becomes feasible to combine genetic and phenotypic data in order to study the immunogenetic architecture of autoimmune disease (Goris and Liston, 2012), much of which is still unknown.

Individuals with one AI disease are at greater risk to develop another AI condition; RA and type 1 diabetes have been observed to occur jointly in patients from a United Kingdom cohort, while there is reduced comorbidity between RA and MS (Somers et al., 2009). Additionally, a GWAS study of inflammatory bowel disease, which includes CD and ulcerative colitis, found variants initially identified in GWAS studies of MS and RA (Liu et al., 2015). In order to delineate potential, common genetic pathways involved in AI and IM disease, we have performed a targeted genetic association study of SNPs known to be associated with AI (MS and RA) and IM (Crohn's disease or CD) in a European American population extracted from the Vanderbilt University Medical Center's (VUMC) DNA biorepository linked to de-identified EHRs. Based on previous evidence from GWAS (Sivakumaran et al., 2011) and candidate gene studies, we hypothesized that MS, RA, and CD share genetic factors within the MHC region, known to include many immune system genes, and outside the MHC region where less is known about genes that may play a role in autoimmunity. Elucidating the role that genetics plays in disease risk and progression is vital for future studies to incorporate risk of a comorbid disorder and lead to better screening and prevention strategies.

## MATERIALS AND METHODS

### Study Population

The study population is derived from BioVU, the VUMC biorepository linked to de-identified EHRs. The Synthetic Derivative or SD refers to the de-identified version of Vanderbilt's EHR and contains inpatient and outpatient medical records from VUMC and affiliated clinics. Patient records consist of both structured (e.g., billing codes, procedure codes, laboratory values) and unstructured (e.g., clinical free text) data. The VUMC EHR contains over 2.2 million records with each record containing 1–1000 International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes captured electronically since the inception of the VUMC EHR more than twenty years ago (Crawford et al., 2015). The SD is linked with VUMC's DNA repository known as BioVU (Roden et al., 2008). These DNA samples are extracted from discarded blood samples collected from outpatient clinical laboratories. Genomic data generated on BioVU samples are available for additional analyses by other investigators after IRB approval. All procedures were approved by the Vanderbilt University's Institutional Review Board that determined that this study met the criteria of "non-human subjects" as no personal identifying information is available to the investigators (The Code of Federal Regulations, 45 CFR 46.102 (f)).

### Phenotype Definitions

Detailed protocols for the identification of MS, RA, and CD phenotypes in this EHR have been previously published (Ritchie et al., 2010). Briefly, cases and controls were defined as follows.

#### Multiple Sclerosis Cases

Cases of MS were extracted from the SD based on the presence of either (1) ICD-9-CM for MS (ICD-9-CM 340) or (2) ICD-9-CM for demyelinating disease of the central nervous system (341.9), transverse myelitis (323.9) AND presence of an MS medication (interferon-beta 1a, interferon-beta 1b, glatiramer, natalizumab, Rituxan) AND a text mention of "multiple sclerosis" AND no mention of an ICD-9-CM for a potentially overlapping autoimmune condition. Due to the complex diagnosis process for

MS, all MS cases were manually reviewed. The MS case algorithm deployed here had a positive predictive value (PPV) of 90%.

### Rheumatoid Arthritis Cases

RA cases contained all of the following: (a) RA ICD-9-CM (714, 714.0, 714.1, or 714.2), (b) RA medication (i.e., methotrexate, sulfasalazine, minocycline, hydroxychloroquine, adalimumab, etanercept, infliximab, gold, azathioprine, rituximab, anakinra, abatacept, leflunomide), (c) text match for "rheumatoid arthritis," and (d) does NOT contain an ICD-9-CM for a potentially overlapping autoimmune condition. A randomized, subset of RA cases ($n = 30$) was manually reviewed for quality assurance. The RA case algorithm deployed here had a PPV of 96.6%.

### Crohn's Disease Cases

CD cases contained both an ICD-9-CM for CD (555.*) and at least one medication for CD (i.e., balsalazide, mesalamine, sulfasalazine, ciprofloxacin, levofloxacin, metronidazole, rifaximin, prednisone, budesonide, azathioprine, mercaptopurine, methotrexate, infliximab, adalimumab, certolizumab, natalizumab). If case records also contained an ICD-9-CM for ulcerative colitis (556.*), the ratio of CD ICD-9-CM codes to ulcerative colitis ICD-9-CM codes had to be greater than two to be considered a case in the study. A randomized, subset of CD cases ($n = 30$) was manually reviewed for quality assurance. The CD case algorithm deployed here had a PPV of 100%.

### Joint Controls

Controls were those individuals whose de-identified EHRs met the control criteria for MS, RA, and CD as previously defined (Ritchie et al., 2010). A randomized, subset of 90 controls was manually reviewed for quality assurance. Control records were *devoid* of all of the following:

a) Text match for "multiple sclerosis," no text mention of a conflicting autoimmune disease for CD or RA.
b) ICD-9-CM: MS ICD-9-CM, any ICD-9-CM for RA, other inflammatory arthritides, and other conflicting autoimmune conditions, inflammatory bowel disease.
c) Medications: MS medications (interferon-beta 1a, interferon-beta 1b, glatiramer, natalizumab, Rituxan).

The PPV of the joint controls was 93.3%.

### Clinical Traits and Variables

Demographic data and laboratory tests were extracted and calculated as follows. For cases, age at diagnosis was defined by the date in the patient's EHR for the first mention of an MS (340, 341.9, 323.9), RA (714, 714.0, 714.1, 714.2), or CD (555.*) ICD-9-CM code. For controls, age was calculated from date of birth as of 2014. Clinical values for body-mass-index (BMI), rheumatoid factor, and C-reactive protein were calculated as the mean of all measurements in a patient's medical record, regardless of case/control status.

## SNP Selection and Statistical Analysis

Targeted genotype data was pulled from BioVU as of February 2014 for samples that were previously genotyped for other research studies. SNPs selected for analysis were associated with MS, RA, or CD in a published GWAS (at $p < 10^{-6}$) identified in the NHGRI GWAS Catalog (Welter et al., 2014) or candidate gene study (at $p < 0.05$) as of 2013 (Supplemental Table 1). In total, genotype data for 345 SNPs in 295 genes or gene regions were accessed and pooled from the following Illumina genotyping arrays: 660W (Denny et al., 2011; Ritchie et al., 2013), 1M (Ng et al., 2014), HumanOmni1-Quad, HumanOmni5-Quad, and Infinium HumanExome BeadChip.

Each SNP considered for analysis was common (minor allele frequency >5%) and passed quality control standards (Hardy Weinberg Equilibrium $p < 0.001$) separately in MS, RA, and CD. Analyses were limited to European American patients, and controls under the age of twenty years were removed from analysis as it cannot be determined if these are true disease-free controls for these later onset conditions. Likewise, cases with age at diagnosis under 20 years were removed from analyses given that the genetic architecture for later onset AI and IM may be different than juvenile onset. Quality control and single SNP tests of association were performed using PLINKv1.07 (Purcell et al., 2007). Each SNP was tested for an association using logistic regression assuming a log-additive genetic model adjusted by age and sex. The Bonferroni significance threshold is set at $1.5 \times 10^{-4}$.

Previous studies have suggested that some autoimmune diseases share a common genetic architecture or associated SNPs have evidence of pleiotropy (Ueda et al., 2003; Criswell et al., 2005; Gough and Simmonds, 2007; Zhernakova et al., 2009; Sivakumaran et al., 2011). To explore whether specific variants are associated with multiple phenotypes in this clinical population, MS, CD, and RA cases were combined into a joint multivariate association test of likelihood ratio for model fit. Ordinal regression adjusted for age and sex was performed on each SNP with MS, RA, and CD modeled as predictors of the genotype. Ordinal regression was performed with the software package MultiPhen (O'Reilly et al., 2012) in R software version 3.2.3 (R Development Core Team, 2013). Twenty-two SNPs from the MultiPhen joint model (i.e., MS, CD, and RA) with a $p < 0.05$ were further characterized using the Pathway Analysis by Randomization incorporating Structure (PARIS), a SNP-based pathway analysis tool that identifies biological pathways significantly enriched by genetic variants adjusted for linkage disequilibrium and gene size (Yaspan et al., 2011).

# RESULTS

## Clinical Population Characteristics

We extracted 129 CD cases, 119 MS cases, 53 RA cases, and 1642 joint controls from the de-identified EHR. Demographics in **Table 1** include information for patients over the age of 20 years for all three phenotypes and joint controls. In general, age and sex distributions were as expected. That is, MS and CD cases were on average younger than controls, while RA cases were on average of a similar age compared with controls. Additionally, MS and RA cases were predominantly female, consistent with the known epidemiology for autoimmune diseases. CD cases were equally likely to be male as female as has been observed in previous

**TABLE 1 | Demographics of the BioVU multiple sclerosis, rheumatoid arthritis, and Crohn's disease cases and controls by sex.**

| | Multiple sclerosis cases | | Rheumatoid arthritis cases | | Crohn's disease cases | | Controls | |
|---|---|---|---|---|---|---|---|---|
| | Male | Female | Male | Female | Male | Female | Male | Female |
| N | 33 | 86 | 17 | 36 | 64 | 65 | 805 | 837 |
| Age in years (SD) | 47.6 (15.2) | 45.4 (12.8) | 60.3 (11.5) | 61.3 (12.4) | 47.5 (15.2) | 49.6 (17.5) | 61.1 (15.3) | 58.3 (16.6) |
| BMI, kg/m$^2$ (SD) | 29.1 (6.2) | 27.3 (6.7) | 30.3 (4.0) | 29.2 (6.7) | 26.7 (6.1) | 26.8 (7.5) | 29.2 (6.7) | 29.4 (12.3) |
| RF, μ/mL (SD) | 26.8 (37.1) | 16.0 (21.1) | 66.5 (60.5) | 190.8 (139.3) | 7.5 (6.4) | 10.5 (7.8) | . | . |
| CRP, mg/L (SD) | 31.4 (50.2) | 37.10 (69.3) | 12.5 (13.8) | 24.9 (53.3) | 42.2 (65.1) | 20.3 (30.8) | . | . |

*Values represent means and standard deviations (SD). Abbreviations: RF, rheumatoid factor; CRP, C-reactive protein.*

epidemiological surveys (Cosnes et al., 2011). Of the MS cases, we were able to manually identify the MS clinical subtype for 48% of patients (Davis et al., 2013; Davis and Haines, 2015). Of these, 42.3% were diagnosed with relapsing-remitting, 33% with secondary progressive, 14% with progressive/relapsing, and 10.2% with primary progressive MS.

## Validation of EHR Phenotype Extraction

To further validate the case/control extraction from the EHR, we performed tests of association for each autoimmune disease to replicate previously-reported associations from the literature. We identified 157, 124, and 64 SNPs previously associated with CD, MS, and RA, mostly from GWAS (Supplemental Table 1). Genotype data for these GWAS-identified variants were available for 120 out of 129, 62 out of 119, and 35 out of 53 cases of CD, MS, and RA identified in the EHR, respectively. At a liberal significance threshold of 0.05, we replicated associations for CD (one), MS (11), and RA (one) (Supplemental Tables 2–4).

Despite differences in sample size, previously reported associations proportionally replicated most often in MS (∼9%) compared with CD (∼1%) and RA (∼1.5%) (**Table 2**). As expected, *HLA-DRA* rs3135388 (De Jager et al., 2009; OR = 3.04; $p = 0.004$) and rs3129889 (Patsopoulos et al., 2011; OR = 3.04; $p = 0.004$), variants in strong linkage disequilibrium with one another, as well as *HLA-DQA1* rs9271366 (Australia New Zealand Multiple Sclerosis Genetics Consortium, 2009; OR = 2.16; $p = 0.003$) were strongly associated with MS in the present study. For CD, we replicated rs7714584 (Franke et al., 2010) located in *IRGM* (OR = 2.66; 95% CI: 1.22-5.78; $p = 0.013$). Additionally, we replicated rs7765379 on chromosome 6 in the RA dataset. SNP rs7765379 was originally identified in a Korean population with an OR of 2.51 (Freudenberg et al., 2011) which is similar to the effect size observed in this study of European Americans (OR = 2.42).

## Evidence for Shared Genetic Architecture

Known autoimmune disease risk loci, such as the MHC and *PTPN22* (Begovich et al., 2004; Bottini et al., 2004), have been associated with multiple autoimmune diseases such as type 1 diabetes, CD, and MS (Wellcome Trust Case Control Consortium, 2007). Three of the most significant associations identified in this MS study, of the total 332 SNPs tested, were originally associated with CD (Barrett et al., 2008; rs2301436; OR = 1.21) and RA (Kochi et al., 2010; Stahl et al., 2010; rs3093024; OR = 1.19 and rs3093023; OR = 1.13; **Table 3**). SNPs rs3093024 and rs3093023 in the chemokine C-C motif receptor 6 (*CCR6*) gene, which is located outside of the MHC region yet plays a role in the immune system via B-cell maturation (Bowman et al., 2000), are located approximately 1500 base pairs apart. At least one of these two SNPs has been associated with RA in European-descent (Stahl et al., 2010), Chinese (Jiang et al., 2014), and Japanese (Kochi et al., 2010) populations with similar effect sizes ranging from an OR of 1.13–1.19. In the present study, the two *CCR6* SNPs were associated with MS but in the opposite direction compared with the published RA studies (**Table 3**).

Eight of the SNPs associated with RA at $p < 0.05$ were originally associated with MS (International Multiple Sclerosis Genetics Consortium, 2011; International Multiple Sclerosis Genetics Consortium et al., 2011; Patsopoulos et al., 2011; rs10866713, rs12644284, rs11962089, rs6952809) and CD (International Multiple Sclerosis Genetics Consortium et al., 2007; Franke et al., 2010; Julià et al., 2014; rs4820425, rs6556412, rs7807268, rs9286879) notably in regions outside of the MHC. The protein expressed by the *TRIM2* gene is part of the tripartite motif family that plays a role in neuroprotection, while deleterious mutations in *TRIM2* are known to cause early-onset axonal neuropathy (Ylikallio et al., 2013). SNP rs12644284 located in *TRIM2* moderately contributed to extreme MS (International Multiple Sclerosis Genetics Consortium, 2011) in a European-descent population (OR = 2.04, allele "G"), and was found to be associated with RA in this study (**Table 3**).

The two most significant associations in this study for CD were initially identified in a meta-analysis of European-descent populations for MS where the opposite alleles for rs1323292 (A) and *MAPK1* rs2283792 (C) contributed to MS risk (OR = 1.12 and OR = 1.10, respectively; International Multiple Sclerosis Genetics Consortium et al., 2011). Lastly, *GCH1* rs3783637(C), associated at $p = 2 \times 10^{-6}$ (OR = 1.10) in a GWAS conducted in a Japanese population of RA (Okada et al., 2012), was associated ($p < 0.05$) with CD in this study for the opposite allele "T."

SNP rs2293370(T) located in an intronic region of *TIMMDC1* on chromosome 3 was associated with both MS [OR = 0.27 (CI 0.09 – 0.80); $p = 0.017$] and CD [OR = 0.25 (CI 0.07 – 0.84); $p = 0.024$] in the same directions in this study. Additionally, SNP rs3781913(A) located within the *PDE2A* gene on chromosome 11 was also found to be associated with RA [OR = 0.32 (CI 0.12 – 0.83); $p = 0.020$] and CD [OR = 0.46 (CI 0.24 – 0.91); $p = 0.024$] in the same direction.

**TABLE 2 | Previously-associated SNPs from Crohn's disease, multiple sclerosis, and rheumatoid arthritis that replicated in the present study of European American patients.**

| SNP | Gene | CHR | CA | OR (95% CI) | $CAF_{cases}$ | $CAF_{control}$ | Disease | *P*-value |
|---|---|---|---|---|---|---|---|---|
| rs7714584 | *IRGM* | 5 | G | 2.66 (1.22–5.78) | 0.19 | 0.09 | CD | 0.013 |
| rs9271366 | *DQA1* | 6 | G | 2.16 (1.30–3.62) | 0.22 | 0.14 | MS | 0.003 |
| rs3135388 | *HLA-DRA* | 6 | T | 3.04 (1.40–6.57) | 0.25 | 0.12 | MS | 0.004 |
| rs3129889 | *HLA-DRB1* | 6 | G | 3.04 (1.40–6.57) | 0.25 | 0.12 | MS | 0.004 |
| rs2243123 | *IL12A* | 3 | C | 2.13 (1.16–3.93) | 0.36 | 0.29 | MS | 0.015 |
| rs2523393 | *HLA-B* | 6 | T | 2.77 (2.50–3.08) | 0.77 | 0.56 | MS | 0.018 |
| rs2546890 | *IL12B* | 5 | A | 1.75 (1.32–2.32) | 0.63 | 0.52 | MS | 0.022 |
| rs12368653 | *AGAP2, CYP27B1* | 12 | A | 1.98 (1.10–3.60) | 0.57 | 0.47 | MS | 0.022 |
| rs4409785 | *RGS14* | 11 | C | 1.81 (1.07–3.09) | 0.20 | 0.16 | MS | 0.027 |
| rs7255066 | *PVR* | 19 | C | 1.64 (1.03–2.60) | 0.39 | 0.28 | MS | 0.035 |
| rs7090512 | *IL2RA* | 10 | C | 1.65 (1.01–2.68) | 0.34 | 0.30 | MS | 0.045 |
| rs7765379 | *HLA-DRB1* | 6 | G | 2.40 (1.13–5.13) | 0.24 | 0.11 | RA | 0.023 |

*Tests of association were performed using logistic regression assuming an additive log model adjusted for age and sex. The threshold for significance was 0.05. For each replicated test of association, the rs number, nearest gene, chromosome (CHR), coded allele (CA), odds ratio (95% confidence interval), coded allele frequency (CAF) for cases and controls, associated disease, and p-value are given.*
*Coded allele (CA).*
*Coded allele frequency (CAF).*
*Odds ratio (OR).*
*95% confidence intervals (CI).*

## Joint Modeling of MS, RA, and CD

By combining cases for MS, RA, and CD in a joint model, we increase power to detect an association while simultaneously testing for variants that correlate with all three phenotypes. In an ordinal regression analysis adjusted for age and sex, 22 SNPs were associated with all three phenotypes at $p < 0.05$ (Supplemental Table 5). Of the ten most significant results (**Table 4**), five were originally identified in CD, three in RA, and two in MS (Supplemental Table 1). Only two of these ten were located within the *HLA* region of chromosome 6 suggesting other genomic regions and pathways are contributing to autoimmunity in this study population.

## Pathway Analysis

Standard approaches for genetic association studies limit small, targeted studies due to power issues and lack of genomic coverage that in turn can hinder the identification of modest main effects and biologically relevant associations. Pathway analysis was performed with results from the joint model in order to increase power and identify pathways which may be driving generalized autoimmunity. Using PARIS, we identified two KEGG pathways that were significantly ($p < 0.05$) enriched for autoimmune disease-associated SNPs identified here. These included the NOD-like receptor signaling pathway ($p = 0.028$; Supplemental Figure 1) and the Shigellosis pathway ($p = 0.015$; Supplemental Figure 2) which contain overlapping gene elements. In healthy humans, the NOD-like receptor signaling pathway plays a role in detecting bacterial pathogens and triggering the innate immune system and localized inflammatory responses (Kanneganti et al., 2007; Shaw et al., 2008). *Shigella*, the primary cause of bacillary dysentery, effectively colonize the human intestines through a type III secretion system that pumps effector proteins into host cells that interfere with the host immune response (Ogawa et al., 2008; Schroeder and Hilbi, 2008). Strong evidence supports that infectious agents may lead to autoimmunity through a number of mechanisms including "epitope spreading" in which a long term immune response to a pathogen can lead to damage of the host cells and release of host antigens being taken up by immune cells and labeled as foreign (Vanderlugt and Miller, 2002; Ercolini and Miller, 2009).

# DISCUSSION

In order to determine whether common autoimmune and immune-mediated diseases share an underlying genetic etiology, we performed a targeted genetic association study of common variants in MS, RA, and CD which had been implicated in previous studies. Additionally, joint modeling of the three phenotypes and pathway analysis was performed which identified overlapping KEGG pathways involved in bacterial pathogenesis. Although this study was underpowered to detect associations with moderate effect sizes below 1.80, we replicated a number of strong, well-studied associations in MS and CD.

Studies of MS have historically been confounded by the clinical heterogeneity of the disease as well as the complex interaction of genetics and environment in disease initiation, progression, and severity (Davis et al., 2013; Davis and Haines, 2015). Despite these obstacles, genetic studies identified the MHC region and the *HLA* genes as major predictors of MS over 40 years ago (1972). The *HLA-DRB1*15* allele is a hallmark susceptibility locus for MS (Caillier et al., 2008), which is marked by rs3135388 (International Multiple Sclerosis Genetics Consortium et al., 2007). Our study replicated this association at $p = 0.003$ with an OR of 3.02 and risk allele frequency (RAF) of 25% in cases and 12% in controls, similar to another publication (OR = 2.7; $RAF_{cases} = 29.8\%$; $RAF_{controls} = 13.6\%$; Briggs et al., 2010).

**TABLE 3 | Common genetic variants associated with multiple autoimmune phenotypes.**

| SNP | Gene | Published phenotype | Current study phenotype | CHR | CA | OR (95% CI) | $CAF_{case}$ | $CAF_{control}$ | *P* |
|---|---|---|---|---|---|---|---|---|---|
| rs2301436 | *FGFR1OP* | CD | MS | 6 | A | 0.38 (0.22–0.63) | 0.38 | 0.49 | 0.0002 |
| rs3093024 | *CCR6* | RA | MS | 6 | A | 0.39 (0.23–0.66) | 0.35 | 0.47 | 0.0004 |
| rs3093023 | *CCR6* | RA | MS | 6 | A | 0.33 (0.16–0.68) | 0.33 | 0.47 | 0.002 |
| rs9271366 | *HLA-DRB1* | MS and CD | MS | 6 | G | 2.16 (1.29–3.62) | 0.22 | 0.14 | 0.003 |
| rs2274910 | *ITLN1* | CD | MS | 1 | T | 0.36 (0.17–0.77) | 0.20 | 0.34 | 0.008 |
| rs2836754 | *RPSAP64-RPL23AP12* | CD | MS | 21 | T | 1.78 (1.13–2.80) | 0.46 | 0.35 | 0.011 |
| rs13126505 | *BANK1* | CD | MS | 4 | A | 3.09 (1.26–7.6) | 0.08 | 0.07 | 0.013 |
| rs4409785 | *CEP57* | MS and RA | MS | 11 | C | 1.81 (1.07–3.09) | 0.20 | 0.16 | 0.027 |
| rs6457617 | *HLA-DQB1* | RA | MS | 6 | T | 0.60 (0.37–0.96) | 0.34 | 0.49 | 0.033 |
| rs13031237 | *REL* | RA | MS | 2 | T | 1.92 (1.05–3.52) | 0.46 | 0.36 | 0.034 |
| rs13017599 | *NONOP2* | RA | MS | 2 | A | 1.92 (1.04–3.53) | 0.46 | 0.36 | 0.036 |
| rs2076756 | *NOD2* | CD | MS | 16 | G | 0.52 (0.28–0.97) | 0.16 | 0.26 | 0.041 |
| rs26232 | *C5 or f30* | RA | MS | 5 | T | 1.98 (1.03–3.82) | 0.39 | 0.31 | 0.042 |
| rs9268853 | *HLA-DRA* | CD/UC | MS | 6 | C | 0.40 (0.16–0.97) | 0.21 | 0.36 | 0.043 |
| rs6448432 | *LOC105374540* | RA | MS | 4 | A | 0.48 (0.24–0.98) | 0.28 | 0.31 | 0.045 |
| rs9286879 | *LOC105371618* | CD | MS | 1 | G | 0.46 (0.21–0.99) | 0.20 | 0.26 | 0.047 |
| rs7765379 | *HLA-DRB1* | CD and RA | RA | 6 | G | 2.4 (1.12–5.12) | 0.24 | 0.11 | 0.023 |
| rs10866713 | *IL12B* | MS | RA | 5 | A | 4.81 (1.20–19.31) | 0.50 | 0.20 | 0.026 |
| rs4820425 | *RBX1* | CD | RA | 22 | A | 2.10 (1.05–4.21) | 0.44 | 0.26 | 0.035 |
| rs6556412 | *IL12B5* | CD | RA | 5 | A | 2.23 (1.04–4.79) | 0.50 | 0.31 | 0.038 |
| rs7807268 | *RNA5SP249-RPL32P17* | CD | RA | 7 | G | 0.12 (0.01–0.93) | 0.10 | 0.49 | 0.043 |
| rs12644284 | *TRIM2* | MS | RA | 4 | G | 0.35 (0.12–0.97) | 0.11 | 0.27 | 0.043 |
| rs6952809 | *CHST12* | MS | RA | 7 | T | 0.37 (0.14–0.98) | 0.15 | 0.32 | 0.047 |
| rs11962089 | *POPDC3* | MS | RA | 6 | G | 3.10 (1.01–9.51) | 0.18 | 0.08 | 0.047 |
| rs1323292 | *RGS1* | MS | CD | 1 | C | 2.80 (1.48–5.27) | 0.36 | 0.16 | 0.001 |
| rs2283792 | *MAPK1* | MS | CD | 22 | T | 2.58 (1.33–5.01) | 0.32 | 0.46 | 0.005 |
| rs3780792 | *VAV2* | MS | CD | 9 | G | 0.61 (0.41–0.92) | 0.26 | 0.34 | 0.018 |
| rs2293370 | *TIMMDC1* | MS | CD | 3 | T | 0.24 (0.07–0.81) | 0.06 | 0.19 | 0.021 |
| rs13119723 | *IL21* | RA | CD | 4 | G | 0.20 (0.05–0.88) | 0.05 | 0.16 | 0.033 |
| rs3781913 | *PDE2A* | RA | CD | 11 | A | 0.48 (0.25–0.95) | 0.26 | 0.43 | 0.036 |
| rs7238078 | *MALT1* | MS | CD | 18 | G | 0.61 (0.39–0.97) | 0.16 | 0.24 | 0.036 |
| rs3783637 | *GCH1* | RA | CD | 14 | T | 1.58 (1.01–2.47) | 0.18 | 0.12 | 0.041 |
| rs2002842 | *LOC105372221* | RA | CD | 18 | A | 1.45 (1.01–2.08) | 0.49 | 0.43 | 0.045 |

*Tests of association were performed using logistic regression assuming a log-additive model adjusted for age and sex. Associations were considered significant at $p < 0.05$. For each significant test of association, rs number, nearest gene, originally-published associated phenotype, the current study phenotype, chromosome (CHR), coded allele (CA), odds ratio (95% confidence interval), coded allele frequency (for cases and controls), and p-value are given.*

*Coded allele for this study (CA).*

*Coded allele frequency (CAF).*

*Odds ratio (OR).*

*95% confidence intervals (CI).*

While we replicated other associations between MS and variants in the *HLA* genes, perhaps more interesting were the inverse associations for SNPs that were identified in studies of RA. These inverse associations may represent false positive associations in the present study (a strong possibility underscoring the small sample sizes and limited power) or interesting complex pleiotropic or co-morbid relationships. SNPs rs3093024 and rs3093023 are both located in the *CCR6* gene on chromosome 6. Originally identified in independent GWAS of RA in a Chinese and Japanese population, these variants contributed to a small increase in risk of RA within these populations (OR = 1.12–1.13; Kochi et al., 2010; Jiang et al., 2014). In the BioVU European American population, these variants were associated with MS in the opposite direction [OR = 0.39–0.43]. Both MS and RA are T helper type 1 cell (Th1)—mediated AI diseases that have inconsistently been observed to co-occur in individuals at a rate varying from the general populace. In a large population study utilizing the United Kingdom General Practice Research Database, an inverse rate of comorbidity was observed between RA and MS (Somers et al., 2009). The opposite was observed in U.S.-based multiplex families with MS in which RA occurred in 2% (Barcellos et al., 2006) of cases vs. ∼1% in the general U.S. population (Helmick et al., 2008). Additionally, a Taiwanese population study of MS

**TABLE 4 | The 10 most significant association results for the ordinal regression joint model adjusted by age and sex.**

| SNP | Gene | Joint Model (*p*-value) |
|---|---|---|
| rs2274910 | *ITLN1* | 0.002 |
| rs6457617 | *HLA-DQB1* | 0.004 |
| rs1551398 | *Chr 8 (intergenic)* | 0.009 |
| rs26232 | *C5 or f30* | 0.010 |
| rs1323292 | *RGS1* | 0.015 |
| rs6651252 | *MIR1208 - MIR3686* | 0.016 |
| rs1800795 | *IL6* | 0.017 |
| rs10734105 | *TCERG1L* | 0.017 |
| rs7765379 | *HLA-DQA2* | 0.019 |
| rs2542151 | *PTPN2* | 0.020 |

and comorbid AI diseases also found patients diagnosed with MS were at greater risk of developing RA (OR = 4.8; Kang et al., 2010). In chart reviews from the Vanderbilt University Medical Center's EHR linked to BioVU, we noted two MS patients out of 162 (1.2%) who were diagnosed with RA. While this rate falls within the range observed in previous studies, it should be noted that the present numbers from BioVU may be biased given that medical charts were not reviewed with the intent to identify additional overlapping AI diseases.

There were a number of factors that limited the strength of this study, particularly the small number of strict cases available with genetic data. Had we accepted a more lenient definition for cases, such as including patients whose medical records contained an ICD-9-CM code for another AI disease, we could have increased the number of cases for RA ($n = 286$). By utilizing a more lenient case definition, we are able to replicate previous RA associations for rs6910071 [(Stahl et al., 2010); OR = 2.41 (CI 1.72–3.39); $p = 3.42 \times 10^{-7}$] and rs660895 [(Plenge et al., 2007); OR = 2.21 (CI 1.60–3.03); $p = 9.98 \times 10^{-7}$] at a Bonferroni significance threshold ($1.5 \times 10^{-4}$).

The strict case definitions also may have limited our ability to more fully describe the shared genetic architecture for these AI and IM outcomes. That is, the case definitions explicitly excluded patients with two or more overlapping conditions (among MS, RA, and CD). Had this exclusion not been in place, the resulting cases would be a mix of bona fide cases with two or more AI or IM conditions and those who had a billing code for one AI or IM disease but were diagnosed with another. It is well known that many of these AI and IM patients experience a diagnostic odyssey. Multiple sclerosis patients, for example, typically receive an eventual differential diagnosis where laboratory tests are ordered to rule out other conditions. More complex data mining followed by manual review of EHRs would be required to identify true cases of AI or IM for study.

Another challenge unique to EHRs compared with epidemiologic cohorts is the patient-to-patient variability related to follow-up data. For example, Davis et al. (2013) noted that the median follow-up time for MS patients in BioVU was 4.5 years, which is similar to the observed median follow-up time for BioVU patients with normal electrocardiograms (5.0 years; Denny et al., 2010) and for BioVU patients with a heart transplant (5.5 years; Oetjens et al., 2014). The range of follow-up on a per patient basis, however, can be extreme (0–20 years for MS patients; Davis et al., 2013). The extreme differences in follow-up time per patient is also reflected on the number of clinic visits available in BioVU per patient. An analysis of ∼15,000 patients in BioVU revealed an average of ∼82 clinic visits per patients with a wide range of one to 1456 clinic visits per patient (Crawford et al., 2015). The paucity of clinical data for a proportion of patients in any given EHR is a known caveat for use of these data in a research setting (Hersh et al., 2013).

For patients with adequate follow-up data, access to the rich, longitudinal EHR data is a strength that enables identification of patients with complex diseases that might be difficult to extract from epidemiological cohorts. Additional work, however, is required to better define clinical populations. Indeed, EHRs in the United States recently adopted the 10th revision of the International Statistical Classification of Diseases and Related Health Problems codes (ICD-10) in response to meaningful use encouraged in part by the Health Information Technology for Economic and Clinical Health (HITECH) Act as part of the American Recovery and Reinvestment Act (ARRA). Compared with ICD-9-CM, the ICD-10 billing codes expanded from 13,000 to 68,000, offering much needed granularity for both clinical care and research relevant to precision medicine.

The most significant individual SNP associations in the joint model were *HLA* variants (**Table 4**) and *ITLN1* (Barrett et al., 2008; Franke et al., 2010; Liu et al., 2015) and *PTPN2* (Okada et al., 2012) genes. PTPN2 is a non-receptor type tyrosine-specific phosphatase that dephosphorylates receptor and non-receptor protein tyrosine kinases in a diverse range of signaling cascades that are responsible for hematopoiesis, cell proliferation, and inflammatory response. PTPN2 also negatively regulates a number of immune pathways responsible for T-cell differentiation (Spalinger et al., 2015) and Interleukin-6 (*IL6;* **Table 4**) cytokine signaling (Yamamoto et al., 2002).

Consistent evidence finds that autoimmune diseases co-occur at an increased rate in family-based and population studies (Barcellos et al., 2006; Eaton et al., 2007; Somers et al., 2009; Sardu et al., 2012) suggesting that these conditions share common genetic or environmental etiologies. Here we implicate bacterial pathogenesis (e.g., NOD-like receptor signaling and Shigellosis) as a common underlying susceptibility factor for MS, CD, and RA. Individual SNPs and genes in these pathways may not be significant at a multiple-correction threshold due to small sample size and clinical heterogeneity as described previously. Yet, pathway analysis takes into consideration modest effect sizes distributed throughout genes in a pathway and increases the likelihood of identifying relevant biological components. Infectious pathogens have been hypothesized to play a major role in the development of autoimmune diseases through a number of mechanisms. A couple of these mechanisms included epitope spreading and molecular mimicry (Ercolini and Miller, 2009). Molecular mimicry results due to a pathogen carrying proteins, amino acid sequence, or antigens that closely resemble host factors which lead to T or B cells triggering an autoimmune response (Lo et al., 2000; Paludan and Bowie, 2013).

In conclusion, pathway analysis of a joint model identified overlapping KEGG pathways with implications for autoimmunity driven by bacterial pathogenesis. These pathways have already been found in individual studies of AI/IM diseases. Future studies with a larger clinical population may allow for replication and fine-tuning of these results. Longitudinal clinical or epidemiological data will be necessary to determine whether acute or chronic infections indeed lead to susceptibility or progression of comorbid AI/IM diseases.

## AUTHOR CONTRIBUTIONS

The listed authors provided substantial contributions to the conception or design of the work (NR and DC), the acquisition (DC, NR, and JM), analysis (NR and MB), or interpretation of the data (NR and DC) for the work; drafted the work (NR) or revised it critically for important intellectual context (NR, MB, JM, and DC); gave final approval of the version to be published (NR, MB, JM, and DC); and agreed to be accountable for all aspects of the work in ensuring that questions related to accuracy or integrity of any part of the work are appropriately investigated and resolved (NR, MB, JM, and DC).

## ACKNOWLEDGMENTS

The dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU which is supported by institutional funding and by the Vanderbilt CTSA grant funded by the National Center for Research Resources, Grant UL1 RR024975-01, which is now at the National Center for Advancing Translational Sciences, Grant 2 UL1 TR000445-06. The Vanderbilt University Center for Human Genetics Research, Computational Genomics Core provided computational and/or analytical support for this work.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2016.00185

## REFERENCES

(1972). Originally published as Volume 2, Issue 7789PRIORITIES IN MEDICINE. *Lancet* 300, 1240. doi: 10.1016/S0140-6736(72)92285-4

(2002). *Autoimmune Diseases Coordinating Committee: Autoimmune Diseases Research Plan*. Report No.: 03-5140, National Institutes of Health. Available online at: http://www.niaid.nih.gov/topics/autoimmune/documents/adccreport.pdf

Australia and New Zealand Multiple Sclerosis Genetics Consortium (2009). (ANZgene) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nat. Genet.* 41, 824–828. doi: 10.1038/ng.396

Barcellos, L. F., Kamdar, B. B., Ramsay, P. P., DeLoa, C., Lincoln, R. R., Caillier, S., et al. (2006). Clustering of autoimmune diseases in families with a high-risk for multiple sclerosis: a descriptive study. *Lancet Neurol.* 5, 924–931. doi: 10.1016/S1474-4422(06)70552-X

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., et al. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* 40, 955–962. doi: 10.1038/ng.175

Begovich, A. B., Carlton, V. E., Honigberg, L. A., Schrodi, S. J., Chokkalingam, A. P., Alexander, H. C., et al. (2004). A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.* 75, 330–337. doi: 10.1086/422827

Bottini, N., Musumeci, L., Alonso, A., Rahmouni, S., Nika, K., Rostamkhani, M., et al. (2004). A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat. Genet.* 36, 337–338. doi: 10.1038/ng1323

Bowman, E. P., Campbell, J. J., Soler, D., Dong, Z., Manlongat, N., Picarella, D., et al. (2000). Developmental switches in chemokine response profiles during B cell differentiation and maturation. *J. Exper. Med.* 191, 1303–1318. doi: 10.1084/jem.191.8.1303

Briggs, F. B., Bartlett, S. E., Goldstein, B. A., Wang, J., McCauley, J. L., Zuvich, R. L., et al. (2010). Evidence for CRHR1 in multiple sclerosis using supervised machine learning and meta-analysis in 12 566 individuals. *Hum. Mol. Genet.* 19, 4286–4295. doi: 10.1093/hmg/ddq328

Caillier, S. J., Briggs, F., Cree, B. A., Baranzini, S. E., Fernandez-Viña, M., Ramsay, P. P., et al. (2008). Uncoupling the roles of HLA-DRB1 and HLA-DRB5 genes in multiple sclerosis. *J. Immunol.* 181, 5473–5480. doi: 10.4049/jimmunol.181.8.5473

Cosnes, J., Gower-Rousseau, C., Seksik, P., and Cortot, A. (2011). Epidemiology and natural history of inflammatory bowel diseases. *Gastroenterology* 140, 1785-1794. doi: 10.1053/j.gastro.2011.01.055

Crawford, D. C., Goodloe, R., Farber-Eger, E., Boston, J., Pendergrass, S. A., Haines, J. L., et al. (2015). Leveraging epidemiologic and clinical collections for genomic studies of complex traits. *Hum. Hered.* 79, 137–146. doi: 10.1159/000381805

Criswell, L. A., Pfieffer, K. A., Lum, R. F., Gonzales, B., Novitzke, J., Kern, M., et al. (2005). Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes. *Am. J. Hum. Genet.* 76, 561–571. doi: 10.1086/429096

Davis, M. F., and Haines, J. L. (2015). The intelligent use and clinical benefits of electronic medical records in multiple sclerosis. *Expert Rev. Clin. Immunol.* 11, 205–211. doi: 10.1586/1744666X.2015.991314

Davis, M. F., Sriram, S., Bush, W. S., Denny, J. C., and Haines, J. L. (2013). Automated extraction of clinical traits of multiple sclerosis in electronic medical records. *J. Am. Med. Informat. Assoc.* 20, e334–e340. doi: 10.1136/amiajnl-2013-001999

De Jager, P. L., Jia, X., Wang, J., de Bakker, P. I. W., Ottoboni, L., Aggarwal, N. T., et al. (2009). Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat. Genet.* 41, 776–782. doi: 10.1038/ng.401

Denny, J. C., Crawford, D. C., Ritchie, M. D., Bielinski, S. J., Basford, M. A., Bradford, Y., et al. (2011). Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am. J. Hum. Genet.* 89, 529–542. doi: 10.1016/j.ajhg.2011.09.008

Denny, J. C., Ritchie, M. D., Crawford, D. C., Schildcrout, J. S., Ramirez, A. H., Pulley, J. M., et al. (2010). Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. *Circulation* 122, 2016–2021. doi: 10.1161/CIRCULATIONAHA.110.948828

Eaton, W. W., Rose, N. R., Kalaydjian, A., Pedersen, M. G., and Mortensen, P. B. (2007). Epidemiology of autoimmune diseases in Denmark. *J. Autoimmun.* 29, 1–9. doi: 10.1016/j.jaut.2007.05.002

Ercolini, A. M., and Miller, S. D. (2009). The role of infections in autoimmune disease. *Clin. Exp. Immunol.* 155, 1–15. doi: 10.1111/j.1365-2249.2008.03834.x

Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* 42, 1118–1125. doi: 10.1038/ng.717

Freudenberg, J., Lee, H.-S., Han, B.-G., Shin, H. D., Kang, Y. M., Sung, Y.-K., et al. (2011). Genome-wide association study of rheumatoid arthritis in Koreans:

population-specific loci as well as overlap with European susceptibility loci. *Arthritis Rheum.* 63, 884–893. doi: 10.1002/art.30235

Gale, A. M. E., and Gillespie, M. K. (2001). Diabetes and gender. *Diabetologia* 44, 3–15. doi: 10.1007/s001250051573

Goris, A., and Liston, A. (2012). The immunogenetic architecture of autoimmune disease. *Cold Spring Harb. Perspect. Biol.* 4:a007260. doi: 10.1101/cshperspect.a007260

Gough, S. C., and Simmonds, M. J. (2007). The HLA rgion and autoimmune disease: associations and mechanisms of action. *Curr. Genomics* 8, 453–465. doi: 10.2174/138920207783591690

Helmick, C. G., Felson, D. T., Lawrence, R. C., Gabriel, S., Hirsch, R., Kwoh, C. K., et al. (2008). Estimates of the prevalence of arthritis and other rheumatic conditions in the United States: Part, I. *Arthritis Rheum.* 58, 15–25. doi: 10.1002/art.23177

Hersh, W. R., Weiner, M. G., Embi, P. J., Logan, J. R., Payne, P. R., Bernstam, E. V., et al. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med. Care* 51(Suppl. 3), S30–S37. doi: 10.1097/MLR.0b013e31829b1dbd

International Multiple Sclerosis Genetics Consortium (2011). Genome-wide association study of severity in multiple sclerosis. *Genes Immun.* 12, 615–625. doi: 10.1038/gene.2011.34

International Multiple Sclerosis Genetics Consortium, Hafler, D. A., Compston, A., Sawcer, S., Lander, E. S., Daly, M. J., et al. (2007). Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* 357, 851–862. doi: 10.1056/NEJMoa073493

International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Control Consortium, Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C. C., et al. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* 476, 214–219. doi: 10.1038/nature10251

Jiang, L., Yin, J., Ye, L., Yang, J., Hemani, G., Liu, A.-J., et al. (2014). Novel risk loci for rheumatoid arthritis in han chinese and congruence with risk variants in Europeans. *Arthritis Rheumatol.* 66, 1121–1132. doi: 10.1002/art.38353

Jin, Y., Birlea, S. A., Fain, P. R., Gowan, K., Riccardi, S. L., Holland, P. J., et al. (2010). Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo. *N. Engl. J. Med.* 362, 1686–1697. doi: 10.1056/NEJMoa0908547

Julià, A., Domènech, E., Chaparro, M., García-Sánchez, V., Gomollón, F., Panés, J., et al. (2014). A genome-wide association study identifies a novel locus at 6q22.1 associated with ulcerative colitis. *Hum. Mol. Genet.* 23, 6927–6934. doi: 10.1093/hmg/ddu398

Kang, J. H., Chen, Y. H., and Lin, H. C. (2010). Comorbidities amongst patients with multiple sclerosis: a population-based controlled study. *Eur. J. Neurol.* 17, 1215–1219. doi: 10.1111/j.1468-1331.2010.02971.x

Kanneganti, T.-D., Lamkanfi, M., and Núñez, G. (2007). Intracellular NOD-like receptors in host defense and disease. *Immunity* 27, 549–559. doi: 10.1016/j.immuni.2007.10.002

Kochi, Y., Okada, Y., Suzuki, A., Ikari, K., Terao, C., Takahashi, A., et al. (2010). A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. *Nat. Genet.* 42, 515–519. doi: 10.1038/ng.583

Laper, S. M., Restrepo, N. A., and Crawford, D. C. (2016). The challenges in using electronic health records for pharmacogenomics and precision medicine research. *Pac. Symp. Biocomput.* 21, 369–380. doi: 10.1142/9789814749411_0034

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47, 979–986. doi: 10.1038/ng.3359

Lo, W.-F., Woods, A. S., DeCloux, A., Cotter, R. J., Metcalf, E. S., and Soloski, M. J. (2000). Molecular mimicry mediated by MHC class Ib molecules after infection with Gram-negative pathogens. *Nat. Med.* 6, 215–218. doi: 10.1038/72329

Ng, M. C. Y., Shriner, D., Chen, B. H., Li, J., Chen, W. M., Guo, X., et al. (2014). Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of Type 2 diabetes. *PLoS Genet.* 10:e1004517. doi: 10.1371/journal.pgen.1004517

Oetjens, M., Bush, W. S., Birdwell, K. A., Dilks, H. H., Bowton, E. A., Denny, J. C., et al. (2014). Utilization of an EMR-biorepository to identify the genetic predictors of calcineurin-inhibitor toxicity in heart transplant recipients. *Pac. Symp. Biocomput.* 2014, 253–264.

Ogawa, M., Handa, Y., Ashida, H., Suzuki, M., and Sasakawa, C. (2008). The versatility of Shigella effectors. *Nat Rev Micro* 6, 11–16. doi: 10.1038/nrmicro1814

Okada, Y., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., et al. (2012). Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. *Nat. Genet.* 44, 511–516. doi: 10.1038/ng.2231

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C. F., Elliott, P., Jarvelin, M. R., et al. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS ONE* 7:e34861. doi: 10.1371/journal.pone.0034861

Paganoni, S., Macklin, E. A., Lee, A., Murphy, A., Chang, J., Zipf, A., et al. (2014). Diagnostic timelines and delays in diagnosing amyotrophic lateral sclerosis (ALS). *Amyotroph. Lateral Scler. Frontotemporal Degener.* 15, 453–456. doi: 10.3109/21678421.2014.903974

Paludan S. R., and Bowie A. G. (2013). Immune sensing of DNA. *Immunity* 38, 870–880. doi: 10.1016/j.immuni.2013.05.004

Patsopoulos, N. A., Bayer Pharma Ms Genetics Working Group, Steering Committees of Studies Evaluating IFNβ-1b and a CCR1-Antagonist, ANZgene Consortium, International Multiple Sclerosis Genetics Consortium, Esposito, F., et al. (2011). Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Ann. Neurol.* 70, 897–912. doi: 10.1002/ana.22609

Plenge, R. M., Seielstad, M., Padyukov, L., Lee, A. T., Remmers, E. F., Ding, B., et al. (2007). TRAF1–C5 as a risk locus for rheumatoid arthritis — a genomewide study. *N. Engl. J. Med.* 357, 1199–1209. doi: 10.1056/NEJMoa073491

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Quan, C., Ren, Y.-Q., Xiang, L.-H., Sun, L.-D., Xu, A.-E., Gao, X.-H., et al. (2010). Genome-wide association study for vitiligo identifies susceptibility loci at 6q27 and the MHC. *Nat. Genet.* 42, 614–618. doi: 10.1038/ng.603

Raychaudhuri, S., Remmers, E. F., Lee, A. T., Hackett, R., Guiducci, C., Burtt, N. P., et al. (2008). Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat. Genet.* 40, 1216–1223. doi: 10.1038/ng.233

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing 2013.

Ritchie, M. D., Denny, J. C., Crawford, D. C., Ramirez, A. H., Weiner, J. B., Pulley, J. M., et al. (2010). Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* 86, 560–572. doi: 10.1016/j.ajhg.2010.03.003

Ritchie, M. D., Denny, J. C., Zuvich, R. L., Crawford, D. C., Schildcrout, J. S., Bastarache, L., et al. (2013). Genome- and phenome-wide analyses of Cardiac conduction identifies markers of arrhythmia risk. *Circulation* 127, 1377–1385. doi: 10.1161/CIRCULATIONAHA.112.000604

Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R., et al. (2008). Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* 84, 362–369. doi: 10.1038/clpt.2008.89

Sardu, C., Cocco, E., Mereu, A., Massa, R., Cuccu, A., Marrosu, M. G., et al. (2012). Population based study of 12 autoimmune diseases in Sardinia, Italy: prevalence and comorbidity. *PLoS ONE* 7:e32487. doi: 10.1371/journal.pone.0032487

Schroeder, G. N., and Hilbi, H. (2008). Molecular pathogenesis of *Shigella* spp.: controlling host cell signaling, invasion, and death by Type III secretion. *Clin. Microbiol. Rev.* 21, 134–156. doi: 10.1128/cmr.00032-07

Shaw, M. H., Reimer, T., Kim, Y.-G., and Nuñez, G. (2008). NOD-like receptors (NLRs): bona fide intracellular microbial sensors. *Curr. Opin. Immunol.* 20, 377–382. doi: 10.1016/j.coi.2008.06.001

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., et al. (2011). Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* 89, 607–618. doi: 10.1016/j.ajhg.2011.10.004

Somers, E. C., Thomas, S. L., Smeeth, L., and Hall, A. J. (2009). Are individuals with an autoimmune disease at higher risk of a second autoimmune disorder? *Am. J. Epidemiol.* 169, 749–755. doi: 10.1093/aje/kwn408

Spalinger, M. R., Kasper, S., Chassard, C., Raselli, T., Frey-Wagner, I., Gottier, C., et al. (2015). PTPN2 controls differentiation of CD4+ T cells and limits intestinal inflammation and intestinal dysbiosis. *Mucosal Immunol.* 8, 918–929. doi: 10.1038/mi.2014.122

Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P., et al. (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* 42, 508–514. doi: 10.1038/ng.582

Ueda, H., Howson, J. M., Esposito, L., Heward, J., Snook, H., Chamberlain, G., et al., (2003). Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* 423, 506–511. doi: 10.1038/nature01621

Vanderlugt, C. L., and Miller, S. D. (2002). Epitope spreading in immune-mediated diseases: implications for immunotherapy. *Nat. Rev. Immunol.* 2, 85–95. doi: 10.1038/nri724

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678. doi: 10.1038/nature05911

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006. doi: 10.1093/nar/gkt1229

Yamamoto, T., Sekine, Y., Kashima, K., Kubota, A., Sato, N., Aoki, N., et al. (2002). The nuclear isoform of protein-tyrosine phosphatase TC-PTP regulates interleukin-6-mediated signaling pathway through STAT3 dephosphorylation. *Biochem. Biophys. Res. Communic.* 297, 811–817. doi: 10.1016/S0006-291X(02)02291-X

Yaspan, B. L., Bush, W. S., Torstenson, E. S., Ma, D., Pericak-Vance, M. A., Ritchie, M. D., et al. (2011). Genetic analysis of biological pathway data through genomic randomization. *Hum. Genet.* 129, 563–571. doi: 10.1007/s00439-011-0956-2

Ylikallio, E., Pöyhönen, R., Zimon, M., De Vriendt, E., Hilander, T., Paetau, A., et al. (2013). Deficiency of the E3 ubiquitin ligase TRIM2 in early-onset axonal neuropathy. *Hum. Mol. Genet.* 22, 2975–2983. doi: 10.1093/hmg/ddt149

Zhernakova, A., van Diemen, C. C., and Wijmenga, C. (2009). Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.* 10, 43–55. doi: 10.1038/nrg2489

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.