ORIGINAL ARTICLE

Phenome-wide association study maps new diseases to the human major histocompatibility complex region

Jixia Liu,[1] Zhan Ye,[2] John G Mayer,[2] Brian A Hoch,[2] Clayton Green,[3] Loren Rolak,[4] Christopher Cold,[5] Seik-Soon Khor,[6] Xiuwen Zheng,[7] Taku Miyagawa,[6,8] Katsushi Tokunaga,[6] Murray H Brilliant,[1] Scott J Hebbring[1]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jmedgenet-2016-103867).

For numbered affiliations see end of article.

Correspondence to
Dr Scott J Hebbring, Marshfield Clinic Research Foundation, Center for Human Genetics, 1000 N Oak Ave, Marshfield, WI 54449, USA; hebbring.scott@mcrf.mfldclin.edu

Received 25 February 2016
Revised 29 April 2016
Accepted 19 May 2016
Published Online First 10 June 2016

ABSTRACT

Background Over 160 disease phenotypes have been mapped to the major histocompatibility complex (MHC) region on chromosome 6 by genome-wide association study (GWAS), suggesting that the MHC region as a whole may be involved in the aetiology of many phenotypes, including unstudied diseases. The phenome-wide association study (PheWAS), a powerful and complementary approach to GWAS, has demonstrated its ability to discover and rediscover genetic associations. The objective of this study is to comprehensively investigate the MHC region by PheWAS to identify new phenotypes mapped to this genetically important region.

Methods In the current study, we systematically explored the MHC region using PheWAS to associate 2692 MHC-linked variants (minor allele frequency ≥0.01) with 6221 phenotypes in a cohort of 7481 subjects from the Marshfield Clinic Personalized Medicine Research Project.

Results Findings showed that expected associations previously identified by GWAS could be identified by PheWAS (eg, psoriasis, ankylosing spondylitis, type I diabetes and coeliac disease) with some having strong cross-phenotype associations potentially driven by pleiotropic effects. Importantly, novel associations with eight diseases not previously assessed by GWAS (eg, lichen planus) were also identified and replicated in an independent population. Many of these associated diseases appear to be immune-related disorders. Further assessment of these diseases in 16 484 Marshfield Clinic twins suggests that some of these diseases, including lichen planus, may have genetic aetiologies.

Conclusions These results demonstrate that the PheWAS approach is a powerful and novel method to discover SNP–disease associations, and is ideal when characterising cross-phenotype associations, and further emphasise the importance of the MHC region in human health and disease.

INTRODUCTION

Over the last decade, more than a thousand unique phenotypes have been associated with thousands of loci by genome-wide association study (GWAS).[1] Interestingly, according to the National Human Genome Research Institute (NHGRI) GWAS Catalog, 2.5% of GWAS SNPs, some with potential pleiotropic properties and 13.5% of phenotypes can be mapped to a 4 Mb region on chromosome 6p, encompassing the major histocompatibility complex (MHC) gene cluster. For example, age-related macular degeneration, drug-induced liver injury and schizophrenia, along with many inflammatory and autoimmune conditions, such as ankylosing spondylitis, psoriasis and type I diabetes, are mapped to the MHC region.[1] In addition to these GWASs, the Immunochip Consortium has fine-mapped approximately 20 autoimmune diseases to the MHC region,[2] while others have studied this region by candidate gene association studies, gene expression studies and protein structural variant analyses.[3–5]

The MHC region is characterised by human leucocyte antigen (*HLA*) class I and class II gene clusters. *HLA* genes encode proteins that modulate both innate and adaptive immune response. HLA class I proteins present epitopes from inside the cell (eg, viruses) to identify cells targeted for cytotoxic T-cell digestion. Class I MHC molecules are presented as transmembrane glycoproteins that consist of two polypeptide chains, α and β2-microglobulins.[6] Class II proteins present foreign antigens from outside the cell to stimulate helper T cells and B cells to activate the complement and antibody systems.[7 8] Class II MHC molecules consist of α and β chains that are encoded by *HLA-DP, HLA-DQ or HLA-DR*.[9] The MHC region is one of the most polymorphic regions in the human genome. Strong genetic associations between the *HLA* variants and autoimmune disease have been established for many years. For example, *HLA-B27* has been known to be the major susceptibility gene for ankylosing spondylitis, a complex disease that is characterised by inflammation and ankylosis. Variants in this gene are present in over 90% of patients with ankylosing spondylitis. In another example, the major T1D susceptibility locus maps to the class II loci *HLA-DRB1* and *HLA-DQB1*. Variants in this region may account for 30%–50% of genetic T1D risk.[10] In addition to *HLA* genes, numerous other genes in the MHC region, such as *TNF, MICA, MICB* and *MOG*, also have apparent immunological roles.[11–13] The potential importance of this genetic locus in disease aetiology is also highlighted by the significant proportion of conditions that are genetically mapped to these genes despite occupying a small proportion (only 0.1%) of the human genome.[1] As such, the phenome-wide association study (PheWAS) design may be a powerful method to evaluate genetic variants in this important genetic region.

To cite: Liu J, Ye Z, Mayer JG, *et al*. *J Med Genet* 2016;**53**:681–689.

The PheWAS approach reverses the paradigm of GWAS by using a genotype-to-phenotype strategy to identify diseases that are associated with an individual genetic variant. A commonality of most PheWASs is the use of an electronic health record (EHR) to define a phenome, often relying on standardised International Classification of Diseases, V.9 (ICD9) codes to define the disease status.[14] A challenge with PheWAS is that some phenotypes may be individually rare. Regardless, the PheWAS approach has demonstrated its capacity to rediscover important genetic associations identified previously by GWAS, has the capacity to identify novel associations and is ideal when characterising cross-phenotype associations.[15–20] This may be particularly relevant for *HLA* variants located in the human MHC region on chromosome 6. For example, the first proof-of-concept PheWAS focused on the *HLA DRB1*1501* variant previously associated with multiple sclerosis (MS). This PheWAS was able to demonstrate the importance of this variant in MS, and was able to identify a novel association with erythematosus conditions,[15] an association that was subsequently confirmed in an independent PheWAS.[17]

With many phenotypes already mapped to the MHC region by GWAS, and with many more phenotypes yet to be studied by GWAS, we hypothesised that this region may contain additional genetic associations and that the PheWAS technique may be leveraged to identify such associations. To address this hypothesis, we conducted a comprehensive PheWAS of 2692 genetic variants across the MHC region spanning 4 Mbs on chromosome 6. PheWAS results confirmed many expected associations and identified many novel associations with immunological diseases not yet assessed by GWAS.

MATERIALS AND METHODS

Ethics statement

This study was approved by the Marshfield Clinic's Institutional Review Board (approval number HEB10112). Written and informed consent was acquired for all participants.

Patient population

Genotyped samples have been described elsewhere[21] and have been applied previously to PheWAS.[14 22 23] Briefly, all genotyped individuals were self-identified white/non-Hispanic Marshfield Clinic patients recruited into the Personalised Medicine Research Project (PMRP). PMRP represents a homogenous population with 77% of participants claiming German ancestry. In this PheWAS, 7481 patients were used for discovery and 3887 patients were used for independent validation. Discovery set participants were all over age 40 (mean 59 years), have on average over 30 years of EHR data and have been genotyped by Illumina HumanCoreExome BeadChip (San Diego, California, USA) as described below. Validation set participants were all over age 50 (mean 74 years) and had comparable years of EHR data.

In addition to PMRP, a cohort of 16 484 Marshfield Clinic twins was evaluated for disease concordance. Diseases studied in this population included eight phenotypes associated with the MHC locus in replication studies. Information used to identify twins included shared last name, date of birth, home address, healthcare/billing account and/or clinical documentation suggesting they were twins. Individuals in Marshfield Clinic Twin Cohort (MCTC) are on average 30 years of age with 8.8 years of EHR data. Although zygosity information is unavailable in MCTC, we developed a method that assesses disease concordance rates in twins to study potential genetic diseases as described previously.[22] Significance was measured by determining if a disease co-occurred in pairs of twins more frequently than by chance, given the disease frequency in the cohort. Like the PheWAS, phenotypes were defined by ICD9 coding.

Genotyping

DNA samples from 7481 patients in the discovery set were genotyped by Illumina HumanCoreExome BeadChip. The exome chip consists of 569 645 variants across the genome. For the purpose of the current study, we analysed SNPs from the MHC region that spans chr6: 29091311-33821793 (hg19). After filtering out poor-quality and rare SNPs , 2692 variants remained with minor allele frequency (MAF) ≥0.01, including 142 previously defined (30 October 2014) 'GWAS significant' SNPs (p≤5.0E-08).[1] For validation, 281 SNPs genotyped on the Illumina Human660W-Quad BeadChip were used.

ICD9 code and phenome

The phenome was defined by patient EHR data as described previously.[17 23 24] Briefly, ICD9 codes were used to define cases and controls at varying levels of phenotypic resolution using a roll-up strategy (eg, ICD9 720.89→720.8*→720.*). Patients coded for any one specific code (eg, ICD9 720.89) became 'cases' for that code, whereas those not coded for the specific code or related codes (eg, ICD9 720*) became 'controls'. For common ICD9 codes (≥300 individuals), cases were defined by those coded two or more times ('rule-of-two'); those coded only once were not considered a case or a control. For less frequent ICD9 codes (<300 individuals), all individuals coded for that ICD9 code were designated as a case. As requested by Marshfield Clinic's Institutional Review Board, case status was not defined for rare ICD9 codes (<9 individuals) to protect patient privacy, as PMRP participants originate from a very specific region in Central Wisconsin. A total of 6221 phenotypes/ICD9 codes defined the phenome for this PheWAS.

Statistics

Because some diseases represented rare phenotypes and some SNPs had low MAFs (<0.05), a Fisher's exact test for allelic association was calculated with Plink V.1.9 (http://pngu.mgh.harvard.edu/purcell/plink/).[25] PheWAS was performed on 2692 SNPs including 142 'GWAS significant' SNPs. A suggestive p value cut-off in the discovery set (p≤5E-05) was used to identify associations to be assessed for independent replication. We further applied logistic regression analysis using sex and years of EHR data as potential covariates for all 6221 phenotypes (see online supplementary tables S1 and S2). The rs1794275 genotype was also included as a potential covariate when further studying lichen planus. False discovery rate (FDR) was calculated.[26] A meta-analysis was conducted using Fisher's method by Ri386 V.3.1.0 (R Development Core Team. R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0. 2011. http://www.R-project.org /) for SNP–disease pairs where phenotypic and genotypic data were available in the independent replication set. This analysis considered the direction of effect (ie, ORs) in the two datasets. *HLA* classical haplotypes were imputed using HIBAG R-package.[27 28] Associations between classical *HLA* haplotypes and lichen planus were analysed using Fisher's exact test that compared each haplotype with all other haplotypes.

To characterise cross-phenotype associations in the MHC region, a trait-based association test (TATES) was conducted for 2692 SNPs, including 142 'GWAS significant' SNPs. TATES

analyses were conducted on ICD9 codes that define 12 broad disease categories and those that mapped to inflammatory phenotypes defined by PheCodes.[29] For individual phenotypes, TATES combines the p values obtained in a single marker test to arrive at a global p value while correcting for observed correlational structure between phenotypes.[20 30]

RESULTS

PheWAS for 'GWAS significant' SNPs

Our initial PheWAS focused on identifying novel associations for GWAS SNPs that mapped to the MHC region [chr6:29091311-33821793 (hg19)]. According to the NHGRI GWAS Catalog, 259 SNPs had 'GWAS significant' associations (p≤5.0E-08);[1] 142 of these SNPs were genotyped by Illumina HumanCoreExome BeadChip. Of these 142 SNPs, 9 were associated with 10 ICD9 codes representing four general phenotypes that passed a conservative Bonferroni correction (p≤5.7E-08; assuming experimentwise α=0.05, 142 SNPs and 6221 phenotypes) in our discovery set (table 1). Three of the nine SNPs agreed with previous GWAS findings. For those that did not agree, most of the phenotypes were not adequately captured in the PheWAS, either because of rarity or lack of specific ICD9 codes that described the expected phenotype. Rediscovered phenotypes included reconfirming associations with coeliac disease, psoriasis and ankylosing spondylitis. For example, we replicated an association between rs2187668 and coeliac disease (p=1.9E-08, FDR=0.0072).[32 33] Rs3131296 and rs2071278, previously shown to be associated with schizophrenia[34] and levels of complement factors C3 and C4,[35] respectively, and in partial linkage disequilibrium (LD) with rs2187668 in the PMRP sample (r^2=0.53 and 0.45, respectively), were also associated with coeliac disease. Similar scenarios of multiple SNPs in LD sharing common associations were observed for other phenotypes (table 1). In addition to these rediscoveries, we also identified an association between rs1794275 and lichen planus (p=1.8E-08, FDR=0.0071). Lichen planus is an inflammatory condition that can affect skin and mucous membranes and has not been studied by GWAS. This SNP was previously shown to be associated by GWAS with IgA nephropathy in an East Asian population.[36] Furthermore, this SNP has been associated with MS, primary biliary cirrhosis, rheumatoid arthritis and type I diabetes by Immunochip Consortium.[2]

PheWAS across entire MHC region

To more comprehensively assess the MHC locus, we conducted PheWAS for all HumanCoreExome SNPs mapped to this region (MAF ≥0.01, n=2692 SNPs). This analysis rediscovered statistically significant associations for psoriasis, ankylosing spondylitis and type I diabetes (p≤3.0E-09; assuming experimentwise α=0.05, 2692 SNPs and 6221 phenotypes). Using a suggestive p value threshold (p≤5.0E-05) in the discovery set, there were 1464 associations consisting of 895 SNPs and 425 phenotypes. Since many of these associations are not overly significant given FDR >0.05, we expect some false positives (see online supplementary table S1). Among these 1464 associations, 470 SNP–disease pairs, consisting of 281 SNPs and 214 phenotypes, could be assessed in the independent validation set with available data. Of the 470 pairs, 64 SNP–disease pairs had suggestive evidence in the validation set (p≤0.05), 58 SNP–disease pairs (91%), consisting of 44 SNPs and 23 phenotypes, demonstrating similar directions of effect in both datasets indicating potential enrichment for true associations (see online supplementary table S2). Among the 23 disease phenotypes, 8 diseases have not been characterised by previous GWAS. These eight diseases were associated with 16 SNPs (table 2). Under this scenario, lichen planus again showed associations with MHC SNPs, including the GWAS SNP described previously (rs1794275). As also mentioned previously and further described below, lichen planus is an immune-related condition. Based on clinical descriptions, many other conditions may also have immunological aetiologies (table 2).

To determine if any of the eight diseases may have underlying genetic aetiologies, 16 484 twins in the MCTC were assessed.[22] The analysis considered the disease concordance rates in families of twins relative to the disease frequency in the population. In general, most diseases were rare in the twin cohort given that MCTC represents younger patients (average 30 years of age) with fewer years of EHR data (average 8.8 years) compared with patients in PMRP (average >59 years of age and >30 years of EHR data). Even with this limitation, three diseases had suggestive evidence (p<0.05) that these conditions cosegregated in families of twins and may be driven in part by genetics. With 46 affected and 1 family of disease-concordant twins, lichen planus approached significance (p=0.062) (see online supplementary table S3). The most significant phenotype included dyshidrosis, a skin condition that results in small fluid-filled blisters often affecting the hands and can be associated with atopic dermatitis and other allergic conditions.[37] In MCTC, 133 twins were affected with dyshidrosis, including seven families of disease-concordant twin families (p=1.6E-6).

Lichen planus associations in the MHC region

Lichen planus is known to be T cell-mediated and can be caused by MHC-linked graft-versus-host disease from allogenic bone marrow transplantation.[38] Since lichen planus was the most significant phenotype identified in the current study, but not previously studied by GWAS, we conducted follow-up analyses. To further assess lichen planus associations in the MHC region, logistic regression was conducted in the discovery set. Except for the association with top 'GWAS significant' SNP rs1794275, lichen planus also showed associations with five other SNPs across the MHC region (p≤1.9E-5; assuming experimentwise α=0.05, 2692 SNPs) (figure 1A). All six SNPs were common variants with similar ORs (2.0–2.5) for the minor alleles. Four of the six associations were confirmed in the validation set (p<0.05) (table 3). However, LD analysis indicated that these SNPs were in partial or strong LD with the most significant SNP (rs1794275), indicating that these SNP associations were likely due to LD. Indeed, significance of these associations was diminished when the effect of rs1794275 was adjusted by logistic regression (figure 1B). For follow-up, and to provide potential functional insights, we assessed associations between classical HLA haplotypes and lichen planus. Haplotype *HLA DQB1*05:01* had the strongest association with lichen planus (p=8.0E-08). Similar haplotype analyses were conducted for the seven other novel phenotypes (table 2 and see online supplementary table S4).

Cross-phenotype association analysis of the MHC region

Because many disease-associated variants have been mapped to the MHC region by GWAS and now PheWAS, we characterised cross-phenotype associations for 2692 HumanCoreExome BeadChip MHC SNPs. Focus was on 358 ICD9 codes that define inflammatory phenotypes.[29] Five SNPs were statistically significant for cross-phenotype associations (p≤1.9E-05; assuming experimentwise α=0.05, 2692 SNPs), including rs4349859, rs4418214, rs9391846, rs12175489 and rs2844505 (figure 2).

Table 1 Significant PheWAS associations (p≤5.7E-08) with 'GWAS significant' SNPs

SNP	Position (bp, hg19)	Gene	Cases (MAF)	Controls (MAF)	OR (95% CI)	Fisher's p value	FDR	ICD9 Code	PheWAS Phenotype	GWAS Phenotype[1]	Reported OR ranges[1]
rs9264942	31 274 380	*HLA-C*, *HLA-B*	327 (0.46)	6744 (0.35)	1.6 (1.4 to 1.9)	5.8E-09	0.0026	696	Psoriasis and similar disorders	Crohn's disease, HIV-1 control	1.2–2.9, 5.3
			253 (0.48)	7034 (0.35)	1.7 (1.4 to 2.0)	5.5E-09	0.0026	696.1	Other psoriasis		
rs10484554	31 274 555	*HLA-C*, *HLA-B*	327 (0.25)	6744 (0.14)	2.0 (1.7 to 2.4)	9.2E-13	1.5E-6	696	Psoriasis and similar disorders	Psoriasis, AIDS progression	2.8–4.7, NR
			42 (0.40)	7376 (0.14)	4.2 (2.7 to 6.4)	1.9E-09	0.001	696.0	Psoriatic arthropathy		
			253 (0.28)	7034 (0.14)	2.4 (1.9 to 2.9)	2.1E-15	5.7E-9	696.1	Other psoriasis		
rs4349859	31 365 787	*HLA-B*, *MICA*	35 (0.39)	7383 (0.04)	14.5 (8.9 to 23.6)	1.5E-19	2.5E-12	720.0	Ankylosing spondylitis	Ankylosing spondylitis	40.8[31]
			180 (0.16)	7238 (0.04)	4.2 (3.2 to 5.7)	8.7E-17	7.3E-10	720	Ankylosing spondylitis and other inflammatory spondylopathies		
			26 (0.40)	7392 (0.04)	15.6 (8.9 to 27.3)	4.1E-16	1.7E-09	720.8	Other inflammatory spondylopathies		
			26 (0.40)	7392 (0.04)	15.6 (8.9 to 27.3)	4.1E-16	1.7E-09	720.89	Other inflammatory spondylopathies		
			30 (0.37)	7388 (0.04)	13.3 (7.8 to 22.7)	9.7E-16	3.2E-09	720.9	Unspecified inflammatory spondylopathy		
rs4418214	31 391 401	*MICA*, *HCP5*	253 (0.16)	7034 (0.08)	2.3 (1.8 to 2.9)	8.0E-10	5.6E-04	696.1	Other psoriasis	HIV-1 susceptibility, HIV-1 control	1.5, 4.4
			35 (0.41)	7383 (0.08)	8.2 (5.1 to 13.2)	1.5E-14	3.6E-08	720.0	Ankylosing spondylitis		
			180 (0.19)	7238 (0.08)	2.6 (2.0 to 3.5)	1.8E-10	1.5E-04	720	Ankylosing spondylitis and other inflammatory spondylopathies		
			26 (0.42)	7392 (0.08)	8.3 (4.6 to 14.7)	1.2E-10	1.1E-04	720.8	Other inflammatory spondylopathies		
			26 (0.42)	7392 (0.08)	8.3 (4.6 to 14.7)	1.2E-10	1.1E-04	720.89	Other inflammatory spondylopathies		
			30 (0.40)	7388 (0.08)	7.7 (4.6 to 13.0)	5.8E-12	8.1E-06	720.9	Unspecified inflammatory spondylopathy		
rs9368699	31 802 541	*C6orf48*, *SNORD48*	253 (0.09)	7034 (0.04)	2.7 (1.9 to 3.6)	3.3E-08	0.011	696.1	Other psoriasis	HIV-1 control	NR
rs2071278	32 165 444	*NOTCH4*	47 (0.40)	7371 (0.16)	3.5 (2.3 to 5.2)	1.8E-08	0.0071	579.0	Coeliac disease	Complement factor C3 and C4 levels	0.1
rs3131296	32 172 993	*NOTCH4*	47 (0.38)	7371 (0.14)	3.8(2.5 to 5.8)	3.3E-09	0.0016	579.0	Coeliac disease	Schizophrenia	1.2
rs2187668	32 605 884	*HLA-DQA1*	47 (0.34)	7371 (0.12)	3.8 (2.4 to 5.8)	1.9E-08	0.0072	579.0	Coeliac disease	Coeliac disease, systemic lupus erythematosus, nephropathy (idiopathic membranous), IgA	6.2–7.0, 2.2, 4.3, 2.5
rs1794275	32 671 248	*HLA-DQB1*, *HLA-DQA2*	97 (0.34)	7321 (0.17)	2.5 (1.8 to 3.4)	1.8E-08	0.0071	697.0	Lichen planus	IgA nephropathy	1.3

Reported OR ranges are from the GWAS Catalog[1] unless otherwise specified.
FDR, false discovery rate; GWAS, genome-wide association study; ICD9, International Classification of Disease V.9; MAF, minor allele frequency; NR, not reported; PheWAS, phenome-wide association study.

Table 2 Top genetic associations with MHC SNPs not previously characterised by GWAS

				Discovery set						Validation set				
Phenotype	ICD9 code	SNP	Position (bp,hg19)	Cases (MAF)	Controls (MAF)	Fisher's p value	OR (95% CI)	FDR	HLA haplotype (p value†)	Cases (MAF)	Controls (MAF)	Fisher's p value	OR (95% CI)	Meta-analysis p value
Unspecified histoplasmosis retinitis	115.92	rs3093983 rs3093978	31 496 925 31 498 497	12 (0.54) 12 (0.54)	7406 (0.18) 7406 (0.18)	5.0E-05 5.0E-05	5.3 (2.4 to 11.8) 5.3 (2.4 to 11.8)	0.57 0.57	*HLA-DQB1* **06:02* (6.7E-4)	8 (0.56) 8 (0.56)	3879 (0.19) 3879 (0.19)	0.0005 0.0005	5.7 (2.1 to 15.2) 5.6 (2.1 to 15.2)	1.2E-07 1.2E-07
Haemangioma of intra-abdominal structures	228.04	rs3131003 rs2523619	31 093 482 31 318 144	31 (0.68) 31 (0.45)	7387 (0.42) 7387 (0.20)	3.7E-05 5.1E-06	2.9 (1.7 to 5.0) 3.3 (2.0 to 5.5)	0.52 0.23	*HLA-C* **12:03* (6.8E-4)	25 (0.62) 25 (0.36)	3862 (0.43) 3862 (0.21)	0.0078 0.011	2.2 (1.2 to 3.9) 2.1 (1.2 to 3.8)	1.2E-06 2.7E-07
Pneumonia due to *Staphylococcus*	482.4	rs3129234 rs3129214 rs756440	33 111 347 33 117 258 33 122 331	15 (0.57) 15 (0.57) 15 (0.57)	7403 (0.23) 7403 (0.23) 7403 (0.23)	4.3E-05 4.4E-05 4.3E-05	4.4 (2.1 to 9.0) 4.4 (2.1 to 9.0) 4.4 (2.1 to 9.0)	0.54 0.54 0.54	*HLA-DPB1* **03:01* (5.3E-4)	6 (0.46) 6 (0.46) 6 (0.46)	3881 (0.24) 3881 (0.24) 3881 (0.24)	0.012 0.022 0.011	2.7 (1.2 to 6.0) 2.7 (1.2 to 6.0) 2.7 (1.2 to 6.0)	2.1E-06 3.9E-06 2.1E-06
Lichen planus	697.0	rs12529049 rs4248166 rs13192471 rs1794275 rs2857106	32 357 715 32 366 421 32 671 103 32 671 248 32 787 570	97 (0.26) 97 (0.30) 97 (0.29) 97 (0.34) 97 (0.33)	7321 (0.15) 7321 (0.18) 7321 (0.14) 7321 (0.17) 7321 (0.20)	4.6E-05 1.4E-05 8.4E-08 1.8E-08 1.4E-05	2.0 (1.5 to 2.8) 2.0 (1.5 to 2.8) 2.5 (1.8 to 3.4) 2.5 (1.8 to 3.4) 2.0 (1.5 to 2.7)	0.55 0.38 0.023 0.0071 0.38	*HLA-DQB1* **05:01* (8.0E-08)	81 (0.22) 81 (0.25) 81 (0.23) 81 (0.23) 81 (0.28)	3806 (0.14) 3806 (0.17) 3806 (0.14) 3806 (0.17) 3806 (0.19)	0.0032 0.013 0.0038 0.031 0.0051	1.8 (1.2 to 2.6) 1.6 (1.1 to 2.3) 1.8 (1.2 to 2.6) 1.5 (1.0 to 2.2) 1.7 (1.2 to 2.4)	6.5E-07 8.3E-07 1.9E-09 3.3E-09 3.4E-07
Dyshidrosis	705.81	rs2844697	30 932 309	374 (0.43)	7044 (0.35)	1.2E-05	1.4 (1.2 to 1.6)	0.34	*HLA-B* **35:01* (0.0030)	146 (0.41)	3741 (0.35)	0.042	1.3 (1.0 to 1.6)	2.1E-06
Other and unspecified nonspecific immunological findings	795.79	rs3094165	29 833 541	215 (0.23)	7012 (0.32)	2.6E-05	0.6 (0.5 to 0.8)	0.47	*HLA-DRB1* **01:03* (0.0012)	127 (0.26)	3668 (0.32)	0.044	0.7 (0.6 to 1.0)	4.5E-06
Infraspinatus (muscle) (tendon) sprain	840.3	rs13198118	30 770 732	10 (0.55)	7408 (0.16)	4.2E-05	6.4 (2.6 to 15.5)	0.53	*HLA-DRB1* **08:10* 0.0030	7 (0.36)	3880 (0.16)	0.034	3.0 (1.0 to 9.0)	5.7E-06
Contusion of wrist	923.21	rs9264942	31 274 380	264 (0.44)	7154 (0.35)	1.9E-05	1.5 (1.2 to 1.8)	0.44	*HLA-C* **03:03* (1.6E-4)	98 (0.46)	3789 (0.36)	0.0076	1.5 (1.1 to 2.0)	6.6E-07

†Reported are the haplotypes with minimum p values.
FDR, false discovery rate; GWAS, genome-wide association study; ICD9, International Classification of Disease V.9; MAF, minor allele frequency; MHC, major histocompatibility complex.

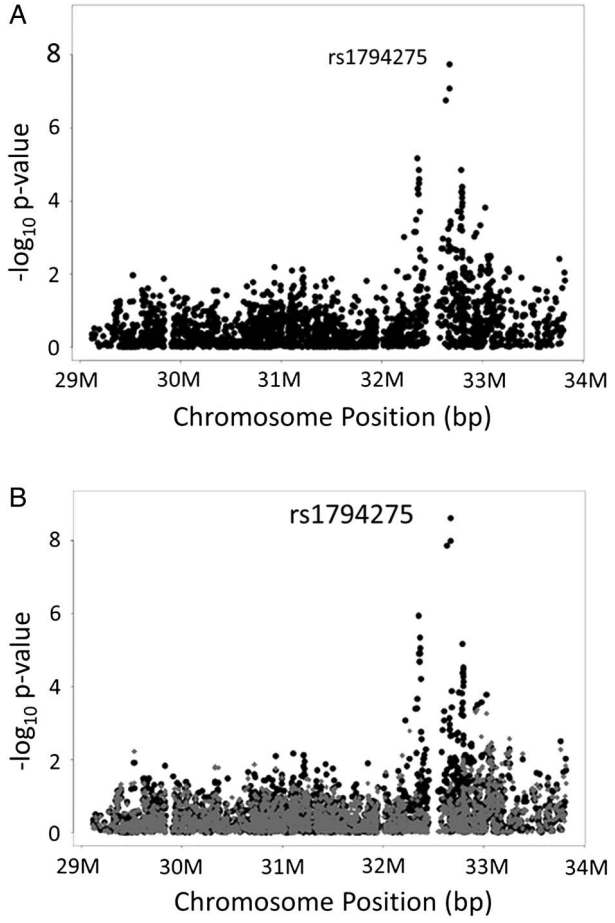

Figure 1 Manhattan plot for lichen planus across the major histocompatibility complex (MHC) region. (A) Fisher's exact analysis and (B) logistic regression analysis. In (B), black data points represent adjustment for gender and years of electronic health record (EHR) data. Grey data points represent adjustment for gender, years of EHR data and rs1794275 genotype.

The top SNP with the most significant cross-phenotype associations was the GWAS SNP rs4349859 (p=2.3E-14). Rs4349859 was most strongly associated with ankylosing spondylitis, as described previously (ICD9 720; p=8.7E-17; table 1), along with multiple arthritic phenotypes including rheumatoid arthritis (ICD9 714; p=0.004), unspecified polyarthropathy or polyarthritis of ankle and foot (ICD9 716.57; p=0.005) and localised osteoarthrosis of the pelvic region and thigh (ICD9 715.35; p=0.008) (see online supplementary table S5). These results are not surprising given the observed genetic overlap between ankylosing spondylitis and rheumatoid arthritis.[39] Finally, rs4349859 was also associated with iridocyclitis (ICD9 364.02; p=8.0E-06) (see online supplementary table S5), an inflammatory condition of the iris and ciliary body that can affect those diagnosed with ankylosing spondylitis.[40] These significant results generated in a single experiment expand on observations of cross-phenotype associations for the MHC region described by many GWASs.

DISCUSSION

Our study is the first comprehensive PheWAS to examine SNPs mapped to the MHC region. As described previously, the MHC region may be involved in the aetiology of many diseases, including unstudied conditions. Diseases such as

Table 3 Top SNP associations with lichen planus

SNP	Position (bp,hg19)	Function	Gene	Amino acid change	Variation	Discovery set					Validation set				Meta-analysis p value
						Case MAF (N=97)	Control MAF (N=7321)	Fisher's p value	OR (95% CI)	FDR	Case MAF (N=81)	Control MAF (N=3806)	Fisher's p value	OR (95% CI)	
rs6930777	32 351 566	Intergenic	*C6orf10, HCG23*	–	C/T	0.23	0.12	6.8E-06	2.3 (1.6 to 3.2)	0.28	–	–	–	–	–
rs4248166	32 366 421	Intronic	*BTNL2*	–	T/C	0.30	0.18	1.4E-05	2.0 (1.5 to 2.8)	0.38	0.25	0.17	0.013	1.6 (1.1 to 2.3)	8.3E-07
rs1049056	32 634 369	Exonic	*HLA-DQB1*	A6S	C/A	0.32	0.16	1.8E-07	2.4 (1.8 to 3.2)	0.03	–	–	–	–	–
rs13192471	32 671 103	Intergenic	*HLA-DQB1, HLA-DQA2*	–	T/C	0.29	0.14	8.4E-08	2.5 (1.8 to 3.4)	0.02	0.23	0.14	0.0038	1.8 (1.2 to 2.6)	1.9E-09
rs1794275	32 671 248	Intergenic	*HLA-DQB1, HLA-DQA2*	–	G/A	0.34	0.17	1.8E-08	2.5 (1.8 to 3.4)	0.007	0.23	0.17	0.031	1.5 (1.0 to 2.2)	3.3E-09
rs2857106	32 787 570	Intergenic	*HLA-DOB, TAP2*	–	T/C	0.33	0.20	1.4E-05	2 (1.5 to 2.7)	0.38	0.28	0.19	0.0051	1.7 (1.2 to 2.4)	3.4E-07

FDR, false discovery rate; MAF, minor allele frequency; N, number of cases or controls.

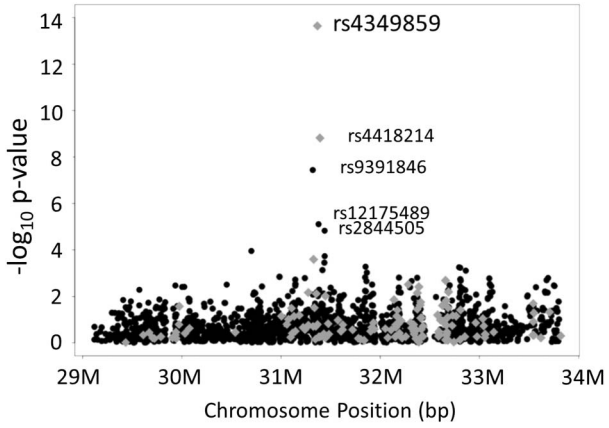

Figure 2 Results of cross-phenotype association analysis of the major histocompatibility complex (MHC) region. Grey data points represent 'genome-wide association study (GWAS) significant' SNPs and black data points represent non-GWAS significant SNPs across the MHC region.

autoimmune, inflammatory and malignant diseases are significantly more common among individuals carrying particular *HLA* alleles.[10] Many of the genes in the MHC region are hypermutable, which is fundamental for their function. This is particularly relevant for *HLA* genes involved in the induction and regulation of the immune responses. As such, the PheWAS may serve as a powerful tool for studying genetic variants in this important genetic locus. Our study confirmed significant associations between SNPs and disease phenotypes identified in previous GWASs,[1] including ankylosing spondylitis, psoriasis, coeliac disease and type I diabetes, some with strong cross-phenotype associations that may be indicative of pleiotropy. Most importantly, we further demonstrated PheWAS' capacity to discover suggestive SNP-disease associations for eight diseases that have not been studied by GWAS. Multiple diseases, including lichen planus, are likely immune-related diseases, reflecting the importance of the MHC region in human immune genetics (see online supplementary table S2). Results from MCTC provide further support that these conditions may be influenced by genetic variation, although we cannot rule out shared environmental effects (see online supplementary table S3). Additional genetic studies focused on these diseases may be needed.

Lichen planus was the most significant disease associated with MHC SNPs not previously studied by GWAS.[1] Our study revealed six SNPs, including one GWAS significant SNP (rs1794275), associated with lichen planus across a 400 kb region, including *HLA-DQB1* (figure 1 and table 3). These six SNPs have also been reported to be associated with MS, type I diabetes and other immune diseases.[2] Interestingly, there is evidence that individuals with MS and type I diabetes are at increased risk for lichen planus.[41 42] Since many of these variants are in partial LD, we suspect that these SNPs probably tag for the same functional variant or haplotype. Haplotype analysis demonstrated *HLA-DQB1*05:01* was strongly associated with lichen planus. Notably, *HLA-DQB1*05:01:01* has been implicated in lichen planus previously by a candidate gene association study.[43] However, other candidate gene studies have implicated other *HLA* haplotypes in lichen planus.[44 45] Since individuals with lichen planus are at increased risk for carcinoma,[46] in-depth future studies of the genetic aetiology of lichen planus may be warranted.

Like GWAS, PheWAS is challenged by multiple comparison testing and frequently applies a Bonferroni adjustment.[14] This study also included variants with MAF ≥0.01, including coding variants, as potential candidates. Variants with MAF <0.05 may have higher effect sizes compared with common variants, but if not, they may increase the burden to identify statistically significant associations. The ability to identify statistical significance is further limited by the inherent nature of the phenotype(s) being studied (eg, heritability, polygenicity, case/control specificity and sample size). In a PheWAS strategy, thousands of phenotypes can be studied simultaneously, but some individual diseases may be rare, have weak genetic aetiologies or be sex-specific. The culmination of these challenges will influence power to detect associations. We attempted to account for these difficulties by applying conservative Fisher's exact tests and Bonferroni adjustments when interpreting results. It is expected that larger cohorts linked to extensive phenotypic data, such as the anticipated 'precision medicine initiative' of over one million individuals,[47] will be ideal for PheWAS to better assess the importance of thousands of phenotypes including rare and sex-specific diseases.

Unique to PheWAS is the potential correlative structure in the phenotypic data. In an ICD9-based PheWAS, similar ICD9 codes may be correlated. Furthermore, correlations exist across codes.[14] Under this circumstance, a stringent Bonferroni correction may be overly conservative. Regardless, we did identify multiple associations that surpassed the conservative multiple testing threshold, but we suspect that additional novel associations remain to be elucidated by follow-up PheWASs and disease-specific studies.

The MHC region brings unique challenges in genetic study not limited to PheWAS. Specifically, this locus includes extreme sequence diversity, structural variants, high gene density with considerable homologies and substantial LD spanning Mbs of sequence.[48] Due to these inherent qualities, characterising the candidate genes and their presumed functional variants can be a significant challenge. These limitations also restrict SNP genotyping array design with many genotyped SNPs, unable to fully address these complexities. It should be noted that other SNP arrays, including Illumina ImmunoChip,[2] provide higher density genotyping for this region compared with the Exome Chip platform. In the future, new long-range sequencing technologies,[49] in combination with the PheWAS approach, may prove necessary to comprehensively study this biologically and clinically important region.

In conclusion, our results expand on what multiple GWASs have identified in that the MHC locus is an important region in human health and disease. Furthermore, this study builds on the growing evidence that the PheWAS technique may be a highly efficient method to detect novel SNP-disease associations and may be ideal when characterising cross-phenotype associations (figure 2). With patient cohorts expanding into the millions linked to extensive phenotypic data, it will be conceivable to conduct a 'GWAS-by-PheWAS' in a single experiment where all diseases are studied at the GWAS level and all variants are studied at the PheWAS level. Such strategies may lead to significant advancements in precision medicine.

Author affiliations
[1]Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, USA
[2]Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, USA
[3]Department of Dermatology, Marshfield Clinic, Marshfield, Wisconsin, USA
[4]Department of Neurology, Marshfield Clinic, Marshfield, Wisconsin, USA

[5]Department of Pathology, Marshfield Clinic, Marshfield, Wisconsin, USA
[6]Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[7]Department of Biostatistics, University of Washington, Seattle, Washington, USA
[8]Sleep Disorders Project, Department of Psychiatry and Behavioral Sciences, Tokyo Metropolitan Institute of Medical Science, Tokyo, Japan

Acknowledgements The authors would like to thank Rachel Stankowski for her assistance in editing this manuscript and Dr Joshua Denny for providing the most recent version of the PheWAS R-package.

Contributors SJH and JL designed the study, and SJH oversaw all aspects of the study. JL performed laboratory experiments. JL, JGM, BAH and ZY generated and analysed association data. CG, LR and CC provided clinical interpretations. S-SK, XZ, TM and KT provided expertise in haplotype analysis. MHB and SJH provided material support. JL and SJH wrote the manuscript with input from all other authors.

Funding This work was supported by the generous donors of the Marshfield Clinic, National Institutes of Health National Center for Advancing Translations Sciences grant number UL1TR000427, National Human Genome Research Institute grant number 1U01HG006389, National Institute of General Medical Science grant number 1R01GM114128 and National Library of Medicine grant number 1K22LM011938.

Competing interests None declared.

Patient consent Obtained.

Ethics approval Marshfield Clinic IRB.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All relevant data are provided in supplemental material. According to Marshfield Clinic IRB, and as approved by USA National Institute of Health, individual-level medical record data cannot be freely shared, to protect patient privacy.

REFERENCES

1 Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, Manolio TA. A Catalog of Published Genome-Wide Association Studies. http://www.genome.gov/gwastudies
2 ImmunoBase. https://www.immunobase.org/page/Welcome/display
3 Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet* 2013;14:661–73.
4 Hanna S, Etzioni A. MHC class I and II deficiencies. *J Allergy Clin Immunol* 2014;134:269–75.
5 Jones EY, Fugger L, Strominger JL, Siebold C. MHC class II proteins and disease: a structural perspective. *Nat Rev Immunol* 2006;6:271–82.
6 Halenius A, Gerke C, Hengel H. Classical and non-classical MHC I molecule manipulation by human cytomegalovirus: so many targets–but how many arrows in the quiver? *Cell Mol Immunol* 2015;12:139–53.
7 Grifoni A, Montesano C, Colizzi V, Amicosante M. Key role of human leukocyte antigen in modulating human immunodeficiency virus progression: an overview of the possible applications. *World J Virol* 2015;4:124–33.
8 Trivedi VB, Dave AP, Dave JM, Patel BC. Human leukocyte antigen and its role in transplantation biology. *Transplant Proc* 2007;39:688–93.
9 Turner D. The human leucocyte antigen (HLA) system. *Vox Sang* 2004;87:87–90.
10 Mosaad YM. Clinical role of human leukocyte antigen in health and disease. *Scand J Immunol* 2015;82:283–306.
11 Silke J, Rickard JA, Gerlic M. The diverse role of RIP kinases in necroptosis and inflammation. *Nat Immunol* 2015;16:689–97.
12 Lanier LL. NKG2D receptor and its ligands in host defense. *Cancer Immunol Res* 2015;3:575–82.
13 Reindl M, Di Pauli F, Rostasy K, Berger T. The spectrum of MOG autoantibody-associated demyelinating diseases. *Nat Rev Neurol* 2013;9:455–61.
14 Hebbring SJ. The challenges, advantages and future of phenome-wide association studies. *Immunology* 2014;141:157–65.
15 Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010;26:1205–10.
16 Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, Chai HS, Bastarache L, Zuvich R, Peissig P, Carrell D, Ramirez AH, Pathak J, Wilke RA, Rasmussen L, Wang X, Pacheco JA, Kho AN, Hayes MG, Weston N, Muatsumoto M, Kopp PA, Newton KM, Karvik GP, Li R, Manolio TA, Kullo IJ, Chute CG, Chisholm RL, Larson EB, McCarty CA, Masys DR, Roden DM, de Andrade M. 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* 2011;89:529–42.
17 Hebbring SJ, Schrodi SJ, Ye Z, Zhou Z, Page D, Brilliant MH. A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun* 2013;14:187–91.
18 Pendergrass SA, Brown-Gentry K, Dudek S, Frase A, Torstenson ES, Goodloe R, Ambite JL, Avery CL, Buyske S, Bůžková P, Deelman E, Fesinmeyer MD, Haiman CA, Heiss G, Hindorff LA, Hsu CN, Jacksin RD, Kooperberg C, Le Marchand L, Lin Y, Matise TC, Monroe KR, Moreland L, Park SL, Reiner A, Wallace R, Wilkens LR, Carwford DC, Ritchie MD. Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* 2013;9:e1003087.
19 Ritchie MD, Denny JC, Zuvich RL, Crawford DC, Schildcrout JS, Bastarache L, Ramirez AH, Mosley JD, Pulley JM, Basford MA, Bradford Y, Rasmussen LV, Pathak J, Chute CG, Kullo IJ, McCarty CA, Chisholm RL, Kho AN, Carlson CS, Larson EB, Jarvik GP, Sotoodehnia N, Cohorts for Heat and Aging Research in Genomic Epidemiology (CHARGE) QRS Group, Manolio TA, Li R, Masys DR, Haines JL, Roden DM. Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 2013;127:1377–85.
20 Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* 2013;14:483–95.
21 McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, Dtruewing JP, Wolf WA, eMERGE Team. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011;4:13.
22 Mayer J, Kitchner T, Ye Z, Zhou Z, He M, Schrodi SJ, Hebbring SJ. Use of an electronic medical record to create the marshfield clinic twin/multiple birth cohort. *Genet Epidemiol* 2014;38:692–8.
23 Ye Z, Mayer J, Ivacic L, Zhou Z, He M, Schrodi SJ, Page D, Brilliant MH, Hebbring SJ. Phenome-wide association studies (PheWASs) for functional variants. *Eur J Hum Genet* 2015;23:523–9.
24 Hebbring SJ, Rastegar-Mojarad M, Ye Z, Mayer J, Jacobson C, Lin S. Application of clinical text data for phenome-wide association studies (PheWASs). *Bioinformatics* 2015;31:1981–7.
25 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. *Am J Hum Genet* 2007;81:559–75.
26 Benjamini YH, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* 1995;57:289–300.
27 Khor SS, Yang W, Kawashima M, Kamitsuji S, Zheng X, Nishida N, Sawai H, Toyoda H, Miyagawa T, Honda M, Kamatani N, Tokunaga K. High-accuracy imputation for HLA class I and II genes based on high-resolution SNP data of population-specific references. *Pharmacogenomics J* 2015;15:530–7.
28 Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, Weir BS. HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenomics J* 2014;14:192–200.
29 Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* 2014;30:2375–6.
30 van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet* 2013;9:e1003235.
31 Evans DM, Spencer CC, Pointon JJ, Su Z, Harvey D, Kochan G, Oppermann U, Dilthey A, Pirinen M, Stone MA, Appleton L, Moutsianas L, Lesli S, Wordsworth T, Kenna TJ, Karaderi T, Thomas GP, Ward MM, Weisman MH, Farrar C, Bradburty LA, Danoy P, Inman RD, Maksymowych W, Gladman D, Rahman P, Spondyloarthritis Research Consortium of Canada (SPARCC), Morgan A, Marzo-Ortega H, Bowness P, Gaddney K, Gaston JS, Smith M, Bruges-Armas J, Couto AR, Sorrentino R, Paladini F, Ferreira MA, Xu H, Liu Y, Jiang L, Kopez-Larrea C, Díaz-Peña R, López-Vázquez A, Zayats T, Band G, Bellenquez C, Blackburn H, Blackwell JM, Bramon E, Bumpstead SK, Casas JP, Corvin A, Craddock N, Deloukas P, Dronov S, Duncanson A, Edkins S, Freeman C, Gillman M, Gray E, Gwilliam R, Hammond N, Hunt SE, Jankowski J, Jayakumar A, Langford C, Liddle J, Markus HS, Mathew CG, McCann OT, McCarthy MI, Palmer CNN, Peltonen L, Plomin R, Potter SC, Rautanen A, Ravindrarajah R, Ricketts M, Samani N, Sawcer SJ, Strange A, Trembath RC, Viswanathan AC, Waller M, Weston P, Whittaker P, Widaa S, Wood NW, McVean G, Reveille JD, Wordsworth BP, Brown MA, Donnelly P, Australo-Angol-American Spondyloarthritis Consortium (TASC); Wellcome Trust Case Control Consortium 2 (WTCCC2). Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat Genet* 2011;43:761–7.
32 Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adány R, Aromaa A, Bardella MT, van den Berg LH, Bockett NA, de la Concha EG, Dema B, Fehrmann RS, Fernández-Arquero M, Fiatal S, Grandone E, Green PM, Groen HJ, Gwilliam R, Houwen RH, Hunt SE, Kaukinen K, Kelleher D, Korponay-Szabo I, Kurppa K, MacMathuna P, Mäki M, Mazzilli MC, McCann OT, Mearin ML, Mein CA, Mirza MM, Mistry V, Mora B, Morley KI, Mulder CJ, Murray JA, Núñez C, Oosterom E, Ophoff RA, Polanco I, Peltonen NL, Platteel M, Rubak A, Salomaa V, Schweizer JJ, Sperandeo MP, Tack GJ, Turner G, Veldink JH, Verbeek WH, Weersma RK, Wolters VM, Urcelay E, Cukrowski B, Greco L, Neuhausen SL, McManus R, Barisani D, Deloukas P, Barrett JC, Saavalainen P, Wijmenga C, van

Heel DA. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* 2010;42:295–302.
33 van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, Barnardo MC, Bethel G, Holmes GK, Feighery C, Jewell D, Kelleher D, Kumar P, Travis S, Walters JR, Sanders DS, Howdle P, Swift J, Playford RJ, McLaren WM, Mearin ML, Mulder CL, McManus R, McGinnis R, Cardon LR, Seloukas P, Wijmenga C. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet* 2007;39:827–9.
34 Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietiläinen OP, Mors O, Mortensen PB, Sigurdsson E, Gustafsson O, Nyegaard M, Tuulio-Henriksson A, Ingason A, Hansen T, Suvisaari J, Lonnqvist J, Paunio T, Børglum AD, Hartmann A, Fink-Jensen A, Nordentoft M, Hougaard D, Norgaard-Pedersen B, Böttcher Y, Olesen J, Breuer R, Möller HJ, Giegling I, Rasmussen HB, Timm S, Matheisen M, Bitter I, Réthelyi JM, Magnusdottir BB, Sigmundsson T, Olason P, Masson G, Gulcher JR, Haraldsson M, Fossdal R, Thorgeirsson TW, Thorsteindottir U, Ruggeri M, Tosato S, Franke B, Strengman E, Kiemeney LA, Genetic Risk and Outcome in Psychosis (GROUP), Melle I, Djurovic S, Abramova L, Kaleda V, Sanjuan J, de Frutos R, Bramon E, Vassos E, Fraser G, Ettinger U, Picchioni M, Walker N, Toulopoulou T, Need AC, Ge D, Yoon JL, Shianna KV, Freimer NB, Cantor RM, Murray R, Kong A, Golimbet V, Carracedo A, Arango C, Costas J, Jönsson EG, Terenius L, Agartz I, Petersson H, Nöthen MM, Rietschel M, Matthews PM, Muglia P, Peltonen L, St Clair D, Godlstein DB, Stefansson K, Collier DA. Common variants conferring risk of schizophrenia. *Nature* 2009;460:744–7.
35 Yang X, Sun J, Gao Y, Tan A, Zhang H, Hu Y, Feng J, Qin X, Tao S, Chen Z, Kim ST, Peng T, Liao M, Lin X, Zhang Z, Tang M, Li L, Mo L, Liang Z, Shi D, Huang Z, Huang X, Liu M, Liu Q, Zhang S, Trent JM, Zheng SL, Xu J, Mo Z. Genome-wide association study for serum complement C3 and C4 levels in healthy Chinese subjects. *PLoS Genet* 2012;8:e1002916.
36 Yu XQ, Li M, Zhang H, Low HQ, Wei X, Wang JQ, Sun LD, Sim KS, Li Y, Foo JN, Wang W, Li ZJ, Yin XY, Tang XQ, Fan L, Chen J, Li RS, Wan JX, Liu ZS, Lou TQ, Zhu L, Huang XJ, Zhang XJ, Liu ZH, Liu JJ. A genome-wide association study in Han Chinese identifies multiple susceptibility loci for IgA nephropathy. *Nat Genet* 2012;44:178–82.
37 Lofgren SM, Warshaw EM. Dyshidrosis: epidemiology, clinical characteristics, and therapy. *Dermatitis* 2006;17:165–81.
38 Nicolatou-Galitis O, Kitra V, Van Vliet-Constantinidou C, Peristeri J, Goussetis E, Petropoulos D, Grafakos S. The oral manifestations of chronic graft-versus-host disease (cGVHD) in paediatric allogeneic bone marrow transplant recipients. *J Oral Pathol Med* 2001;30:148–53.
39 Jawaheer D, Seldin MF, Amos CI, Chen WV, Shigeta R, Monteiro J, Kern M, Criswell LA, Albani S, Nelson JL, Clegg DO, Pope R, Schroeder HW Jr, Bridges SL Jr, Pisetsky DS, Ward R, Kastner DL, Wilder RL, Pincus T, Callahan LF, Flemming D, Wener MH, Gregersen PK. A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. *Am J Hum Genet* 2001;68:927–36.
40 Chung YM, Yeh TS, Liu JH. Clinical manifestations of HLA-B27-positive acute anterior uveitis in Chinese. *Zhonghua Yi Xue Za Zhi (Taipei)* 1989;43:97–104.
41 Sepic J, Ristic S, Perkovic O, Brinar V, Lipozencic J, Crnic-Martinovic M, Starcevic Cizmarevic N, Janko Labinac D, Kapovic M, Peterlin B. A case of lichen ruber planus in a patient with familial multiple sclerosis. *J Int Med Res* 2010;38:1856–60.
42 Mohsin SF, Ahmed SA, Fawwad A, Basit A. Prevalence of oral mucosal alterations in type 2 diabetes mellitus patients attending a diabetic center. *Pak J Med Sci* 2014;30:716–19.
43 Setterfield JF, Neill S, Shirlaw PJ, Theron J, Vaughan R, Escudier M, Challacombe SJ, Black MM. The vulvovaginal gingival syndrome: a severe subgroup of lichen planus with characteristic clinical features and a novel association with the class II HLA DQB1*0201 allele. *J Am Acad Dermatol* 2006;55:98–113.
44 Pavlovsky L, Israeli M, Sagy E, Berg AL, David M, Shemer A, Klein T, Hodak E. Lichen planopilaris is associated with HLA DRB1*11 and DQB1*03 alleles. *Acta Derm Venereol* 2015;95:177–80.
45 Gao XH, Barnardo MC, Winsey S, Ahmad T, Cook J, Agudelo JD, Zhai N, Powell JJ, Fuggle SV, Wojnarowska F. The association between HLA DR, DQ antigens, and vulval lichen sclerosus in the UK: HLA DRB112 and its associated DRB112/DQB10301/04/09/010 haplotype confers susceptibility to vulval lichen sclerosus, and HLA DRB10301/04 and its associated DRB10301/04/DQB10201/02/03 haplotype protects from vulval lichen sclerosus. *J Invest Dermatol* 2005;125:895–9.
46 Gandolfo S, Richiardi L, Carrozzo M, Broccoletti R, Carbone M, Pagano M, Vestita C, Rosso S, Merletti F. Risk of oral squamous cell carcinoma in 402 patients with oral lichen planus: a follow-up study in an Italian population. *Oral Oncol* 2004;40:77–83.
47 Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;372:793–5.
48 Allcock RJ, Atrazhev AM, Beck S, de Jong PJ, Elliott JF, Forbes S, Halls K, Horton R, Osoegawa K, Rogers J, Sawcer S, Todd JA, Trowsdale J, Wang Y, Williams S. The MHC haplotype project: a resource for HLA-linked association studies. *Tissue Antigens* 2002;59:520–1.
49 Ammar R, Paton TA, Torti D, Shlien A, Bader GD. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Res* 2015;4:17.