

## ARTICLE OPEN

## Exome-wide study of ankylosing spondylitis demonstrates additional shared genetic background with inflammatory bowel disease

Philip C Robinson<sup>1,2,3</sup>, Paul J Leo<sup>4</sup>, Jennifer J Pointon<sup>5</sup>, Jessica Harris<sup>4</sup>, Katie Cremin<sup>2</sup>, Linda A Bradbury<sup>4</sup>, Simon Stebbings<sup>6</sup>, Andrew A Harrison<sup>7</sup>, Australian Osteoporosis Genetics Consortium<sup>10</sup>, Wellcome Trust Case Control Consortium<sup>10</sup>, Emma L Duncan<sup>4,8,9</sup>, David M Evans<sup>2</sup>, Paul B Wordsworth<sup>3</sup> and Matthew A Brown<sup>4</sup>

Ankylosing spondylitis (AS) is a common chronic immune-mediated arthropathy affecting primarily the spine and pelvis. The condition is strongly associated with *HLA-B\*27* as well as other human leukocyte antigen variants and at least 47 individual non-MHC-associated variants. However, substantial additional heritability remains as yet unexplained. To identify further genetic variants associated with the disease, we undertook an association study of AS in 5,040 patients and 21,133 healthy controls using the Illumina Exomechip microarray. A novel association achieving genome-wide significance was noted at *CDKAL1*. Suggestive associations were demonstrated with common variants in *FAM118A*, *C7orf72* and *FAM114A1* and with a low-frequency variant in *PNPLA1*. Two of the variants have been previously associated with inflammatory bowel disease (IBD; *CDKAL1* and *C7orf72*). These findings further increase the evidence for the marked similarity of genetic risk factors for IBD and AS, consistent with the two diseases having similar aetiopathogenesis.

npj Genomic Medicine (2016) 1, 16008; doi:10.1038/npjgenmed.2016.8; published online 4 May 2016

## INTRODUCTION

Ankylosing spondylitis (AS) is an inflammatory arthropathy with a prevalence of 0.1–0.5% in populations of European or Asian descent. The condition primarily affects the sacroiliac joints and the spine, initially causing pain and reversible stiffness. Subsequent ankylosis of these leads to fixed spinal deformity and increasing disability. The condition also has extra-articular manifestations, most commonly in the eye (acute anterior uveitis), and rarely in the aorta, kidneys and lungs. Inflammatory bowel disease (IBD; either Crohn's disease or ulcerative colitis) is found in 5–10% of AS patients, and ~60% of AS patients have subclinical ileal inflammation. The disease is known to be highly heritable, and to date at least 47 independent genetic variants have been shown to be associated with AS susceptibility.<sup>1</sup> These findings have contributed greatly to increased understanding of the pathogenesis of the disease, and also to the development of new treatments.<sup>2</sup> The genetic variants associated with AS have been discovered using coding variant scans,<sup>3</sup> genome-wide association studies,<sup>4,5</sup> candidate gene studies<sup>6–10</sup> and studies targeting immunologically important regions.<sup>11</sup> Generally these associations have been with common variants. For example, in the International Genetics of AS Consortium ImmunoChip study, which identified 37 AS-associated variants, the mean minor

allele frequency (MAF) of the key-associated single-nucleotide polymorphism (SNP) was 32%, and only two low-frequency (MAF 1–5%) AS-associated alleles were identified and one rare allele (MAF < 1%).<sup>11</sup> As with many common heritable diseases, a large proportion of the heritability of AS remains unexplained. The vast majority of human genetic variation consists of low-frequency and rare variants with MAF < 5%.<sup>12</sup> Studies performed to date have not addressed such variants extensively due to the design of genome-wide association study SNP microarrays and the sample sizes of the studies not having power to detect association with low-frequency or rare variants.<sup>13</sup> It is likely that there is a substantial numbers of additional variants in the low-frequency or rare variant range that remain undiscovered.

Exomes are proposed to carry a disproportionately high number of clinically important variants because of their potentially profound effects on the protein function.<sup>14</sup> Much of the variation found in exomes is also rare.<sup>15,16</sup> Recent exome-sequencing analyses have demonstrated the huge number of rare and potentially damaging variants present in human exomes, with ~313 genes per person predicted to be adversely affected by exonic variants.<sup>14</sup>

In this study, we sought to identify further AS-associated genetic variants focusing on exons using the Illumina Human-Exome Beadchip microarray. In addition to covering common

<sup>1</sup>School of Medicine, Faculty of Medicine and Biomedical Sciences, University of Queensland, Brisbane, QLD, Australia; <sup>2</sup>University of Queensland Diamantina Institute, Translational Research Institute, Princess Alexandra Hospital, Brisbane, QLD, Australia; <sup>3</sup>Department of Rheumatology, Royal Brisbane and Women's Hospital, Herston, QLD, Australia; <sup>4</sup>Institute of Health and Biomedical Innovation, Queensland University of Technology, Translational Research Institute, Princess Alexandra Hospital Herston, QLD, Australia; <sup>5</sup>National Institute for Health Research (NIHR) Oxford Comprehensive Biomedical Research Centre and Musculoskeletal Biomedical Research Unit, Nuffield Orthopaedic Centre, Oxford, UK; <sup>6</sup>Department of Medicine, University of Otago, Dunedin, New Zealand; <sup>7</sup>Department of Medicine, University of Otago, Wellington, New Zealand; <sup>8</sup>Department of Endocrinology and Diabetes, Royal Brisbane and Women's Hospital, Herston, QLD, Australia and <sup>9</sup>School of Medicine and University of Queensland Centre for Clinical Research, Faculty of Medicine and Biomedical Sciences, University of Queensland, Brisbane, QLD, Australia.

Correspondence: MA Brown (matt.brown@qut.edu.au)

<sup>10</sup>The members of the two Consortia are listed after references.

Received 11 February 2016; accepted 28 March 2016

coding variants, this chip has extensive low-frequency and rare variant content, enabling us to perform a relatively low cost survey of the role of such variants in AS (compared with studies utilising whole-exome or -genome sequencing).

## RESULTS

### Participants and SNPs

After participant quality control (QC) there were 4,602 AS cases and 20,164 healthy controls. After SNP QC there were 207,193 SNPs. Principal components analysis was performed with 0–10 eigenvectors; the scree plot of genomic control-1000 statistics for different numbers of eigenvectors is shown in Supplementary Figure 1. Use of one eigenvector produced a genomic control-1000 statistic of 1.045, and use of additional eigenvectors did not reduce the genomic control-1000 statistic further. Therefore, all further analyses used a single eigenvector to control for population stratification. The quantile–quantile plot for the experiment is shown in Supplementary Figure 2.

### Single variant analysis

Assays for 10 non-MHC SNPs previously associated with AS were included on the HumanExome Beadchip, and each one showed at least suggestive association ( $P < 5 \times 10^{-6}$ ) with AS in the current study (Table 1 and Supplementary Figures 3–13). The associations with SNPs in *ERAP1*, *IL23R*, the intergenic regions chromosomes 2p15 and 21q22, *GPR35* and *IL6R* were confirmed at genome-wide significance level ( $P < 5 \times 10^{-8}$ ). SNP associations within *ANTXR2*, *FCGR2A*, *IL1R1* and *NOS2* were confirmed at suggestive levels of significance. At nine of the confirmed loci, association was seen with the primary associated SNP. At *NOS2* the association was with rs2297518, which has been reported previously as a secondary signal at the *NOS2* locus, the primary associated SNP rs2531875<sup>11</sup> not being included on the HumanExome Beadchip.

Novel common variant association was identified in *CDKAL1* (MAF=0.22,  $P=1.8 \times 10^{-8}$ , OR=1.2) at a genome-wide level of significance (Table 2 and Supplementary Figure 14). The *CDKAL1* SNP rs6908425 has previously been associated with Crohn's disease,<sup>17,18</sup> psoriasis<sup>19</sup> and ulcerative colitis<sup>20</sup> again with the C allele being the risk variant. rs6908425 is in strong linkage disequilibrium with variants associated with diabetes and body mass index (Supplementary Table 1).<sup>21</sup>

Suggestive associations were seen with common variants in *FAM118A* ( $P=5.9 \times 10^{-8}$ , odds ratio (OR)=1.2), *C7orf72* ( $P=1.9 \times 10^{-7}$ , OR=1.1) and *FAM114A1* ( $P=1.4 \times 10^{-6}$ , OR=1.1), and with a rare variant in patatin-like phospholipase domain containing 1 (*PNPLA1*) (MAF=0.0017,  $P=1.5 \times 10^{-6}$ , OR=2.6). SNPs in *C7orf72*, including the same SNP we have identified as being associated with AS (rs1456896), have been associated with Crohn's disease.<sup>18</sup> *FAM118A* variants have previously been associated with bone mineral density.<sup>22</sup> Strong linkage disequilibrium exists between the bone mineral density-associated SNP (rs136564) and the AS-associated SNP, rs6007594 ( $r^2=0.03$ ,  $D'=0.92$ ). Functional annotation for the variants is shown in Supplementary Table 2.

After conditioning on the primary association signals there were second, independent signals observed in *ERAP1* and *IL23R* (Table 3 and Supplementary Figure 19). The pattern of association at *ERAP1* is consistent with previous work that describes two independent haplotypes at *ERAP1*, tagged by the functional variants rs30187-T (primary signal) and rs10050860-C (primary unconditioned signal,  $P=3.1 \times 10^{-24}$ , allele C, OR=1.4, 95% confidence interval (CI) 1.27–1.43 and conditioned on rs30187  $P=9.7 \times 10^{-10}$ , allele C, OR=1.2, 95% CI 1.14–1.30). Moderate linkage disequilibrium was observed between rs30187 and rs10050860 ( $r^2=0.156$  and  $D'=1.00$ ). As previously reported, after conditioning on rs30187, association was confirmed with the neighbouring *ERAP2* gene (rs2549794,  $P=2.8 \times 10^{-14}$ , allele C, OR=1.2, 95% CI 1.16–1.28). This association remains after controlling for both the rs30187 and rs10050860 *ERAP1* variants (rs2549794,  $P=4.7 \times 10^{-14}$ , allele C, OR=1.2, 95% CI 1.16–1.28). No additional associated variants were evident.

At *IL23R*, the intronic variant rs11465804 was slightly more significantly associated than the previously reported rs11209026 nonsynonymous SNP ( $P=1.3 \times 10^{-25}$  and  $P=4.55 \times 10^{-25}$ , respectively). These two SNPs are in high linkage disequilibrium ( $r^2=0.89$ ,  $D'=0.97$ ), and conditioning for either SNP controlled for the association at the other SNP, suggesting that the rs11465804 variant tags this previously demonstrated functional SNP.<sup>23,24</sup> As previously reported, a secondary association is seen at the *IL23R* locus, evident after conditioning on rs11465804 (rs10889677,  $P=9.8 \times 10^{-9}$ , allele A, OR=1.2, 95% CI 1.10–1.20).<sup>3</sup> No additional associated variants were evident.

An independent signal supported by multiple SNPs was demonstrated after conditioning on the main SNP at the *FAM114A1* locus; these SNPs sit in the *TLR10* gene (Table 3 and

**Table 1.** Existing AS genetic associations and new associations in previously associated loci

Gene/locus	Cs	Pos <sup>a</sup>	Current SNP	P-value	Risk/Prot	OR, 95% CI	RAF case/Con	Previously reported SNP	r <sup>2</sup>	D'	Discovery Study
<i>ERAP1</i>	5	96,244,549	rs30187	$3.0 \times 10^{-34}$	T/C	1.3 (1.28–1.41)	0.40/0.34	rs31087	1.00	1.00	Burton et al. <sup>3</sup>
<i>IL23R</i>	1	67,702,526	rs11465804	$1.3 \times 10^{-25}$	T/G	1.9 (1.65–2.01)	0.96/0.93	rs11209026	0.89 <sup>b</sup>	0.97 <sup>b</sup>	Burton et al. <sup>3</sup>
<i>IL23R</i>	1	67,702,526	rs11465804	$1.3 \times 10^{-25}$	T/G	1.9 (1.65–2.01)	0.96/0.93	rs11209032	0.03	1.00	Evans et al. <sup>4</sup>
2p15	2	62,551,472	rs10865331	$1.2 \times 10^{-22}$	A/G	1.3 (1.20–1.32)	0.43/0.38	rs10865331	1.00	1.00	Reveille et al. <sup>5</sup>
21q22	21	40,465,534	rs2836878	$1.1 \times 10^{-16}$	G/A	1.3 (1.19–1.32)	0.77/0.73	rs2836878	1.00	1.00	Reveille et al. <sup>5</sup>
<i>KIF21B-GPR25</i>	1	200,877,562	rs7554511	$3.8 \times 10^{-16}$	C/A	1.2 (1.18–1.30)	0.75/0.71	rs2297909	0.89	0.96	Evans et al. <sup>4</sup>
<i>GPR35</i>	2	241,569,692	rs3749171	$5.0 \times 10^{-12}$	T/C	1.2 (1.15–1.29)	0.21/0.18	rs4676410	0.86	1.00	Cortes et al. <sup>11</sup>
<i>IL6R</i>	1	154,426,970	rs2228145	$1.9 \times 10^{-8}$	A/C	1.1 (1.09–1.20)	0.62/0.59	rs4129267	0.97	1.00	Cortes et al. <sup>11</sup>
<i>ANTXR2</i>	4	80,949,829	rs4333130	$5.3 \times 10^{-8}$	T/C	1.1 (1.09–1.20)	0.66/0.63	rs4333130	1.00	1.00	Evans et al. <sup>4</sup>
<i>IL1R1</i>	2	102,663,628	rs2310173	$2.3 \times 10^{-7}$	T/G	1.1 (1.08–1.18)	0.50/0.47	rs4851529	0.48	0.89	Cortes et al. <sup>11</sup>
<i>IL1R1</i>	2	102,663,628	rs2310173	$2.3 \times 10^{-7}$	T/G	1.1 (1.08–1.18)	0.50/0.47	rs2192752	0.02	0.23	Cortes et al. <sup>11</sup>
<i>FCGR2A</i>	1	161,479,745	rs1801274	$4.9 \times 10^{-6}$	A/G	1.1 (1.06–1.16)	0.50/0.47	rs1801274	1.00	1.00	Cortes et al. <sup>11</sup>
<i>NOS2</i>	17	26,096,597	rs2297518	$2.7 \times 10^{-6}$	A/G	1.1 (1.08–1.21)	0.21/0.19	rs2297518	1.00	1.00	Cortes et al. <sup>11</sup>

Abbreviations: AS, ankylosing spondylitis; CI, confidence interval; Cs, chromosome; OR, odds ratio; Prot, protective; Pos, position; RAF, risk allele frequency; SNP, single-nucleotide polymorphism.

Linkage disequilibrium calculated from 1000 Genomes data via the SNAP Browser from the Broad Institute.

<sup>a</sup>Human Genome Build 19.

<sup>b</sup>Calculated from this data set.

**Table 2.** Novel AS associations

Chr	SNP	Position <sup>a</sup>	Risk allele	Prot allele	Cases RAF	Cont RAF	OR (95% CI)	P-value	Gene/nearby gene
<i>Genome-wide-significant association</i>									
6	rs6908425	20,728,731	C	T	0.81	0.78	1.2 (1.11–1.25)	$1.8 \times 10^{-8}$	<i>CDKAL1</i>
<i>Suggestive association</i>									
22	rs6007594	45,728,370	G	A	0.77	0.74	1.2 (1.10–1.22)	$5.9 \times 10^{-8}$	<i>FAM118A</i>
7	rs1456896	50,304,461	T	C	0.69	0.66	1.1 (1.08–1.20)	$1.9 \times 10^{-7}$	<i>C7orf72</i>
4	rs11555334	38,880,046	T	C	0.75	0.72	1.1 (1.08–1.20)	$1.4 \times 10^{-6}$	<i>FAM114A1</i>
6	rs141744967	36,270,205	T	C	0.0044	0.0017	2.6 (1.77–3.89)	$1.5 \times 10^{-6}$	<i>PNPLA1</i>

Abbreviations: AS, ankylosing spondylitis; Chr, chromosome; CI, confidence interval; Cont, control; OR, odds ratio; Prot, protective; RAF, risk allele frequency.  
<sup>a</sup>Human Genome build 19.

**Table 3.** Secondary associations observed at loci associated with AS at genome-wide significance in the current study

Chr	SNP	Position <sup>a</sup>	Gene	P-value	Risk/non-risk	RAF case/Con	OR (95% CI)	Conditional SNP
5	rs2549794	96,244,549	<i>ERAP2</i>	$2.9 \times 10^{-14}$	C/T	0.44/0.43	1.22 (1.16–1.28)	rs30187
5	rs10050860	96,122,210	<i>ERAP1</i>	$1.6 \times 10^{-9}$	C/T	0.82/0.77	1.21 (1.14–1.29)	rs30187/rs2549794
1	rs10889677	67,725,120	<i>IL23R</i>	$9.7 \times 10^{-9}$	A/C	0.69/0.65	1.15 (1.10–1.21)	rs11465804
4	rs11096955	38,776,107	<i>TLR10</i>	$3.5 \times 10^{-4}$	G/T	0.32/0.30	1.09 (1.04–1.15)	rs11555334

Abbreviations: OR, odds ratio; Chr, chromosome; CI, confidence interval; RAF, risk allele frequency; SNP, single-nucleotide polymorphism.  
<sup>a</sup>Human Genome build 19.

Supplementary Figure 22). SNPs in *TLR10* have previously been associated with Crohn's disease,<sup>25,26</sup> and the peak-associated SNP in our study, rs1109695, is in strong linkage disequilibrium with the most strongly associated, previously reported, Crohn's disease *TLR10* SNPs, rs7658893 ( $r^2 = 0.64$ ,  $D' = 1.0$ ) and rs6841698 ( $r^2 = 0.64$ ,  $D' = 0.83$ ).

#### Rare variant burden tests

Using the SKAT-O test, no genes were significantly associated with AS at either a genome-wide or suggestive level of significance. This was the case for all the sets of variants incorporated into the burden equation including the following: (a) 'Damaging set', (b) 'Polyphen set' and (c) all variants. No association was identified even after removing the restrictive minor allele filter of 0.05 and including SNPs of all frequencies.

With sample sizes of 4,602 AS cases and 20,164 controls used in this study, the power to detect the association was good for low MAF variants, but low for rare variants (Supplementary Figure 23). Assuming a prevalence of 0.55%, allele risk of 1.5 and alpha of  $5 \times 10^{-8}$ , and equal MAF for the disease-causative and genotyped markers, the study had 100% power for MAF = 0.05, but only 9% power for MAF = 0.01 and close to zero ( $6.1 \times 10^{-5}$ ) power for the median MAF in this study ( $2.0 \times 10^{-4}$ ).

## DISCUSSION

This study re-demonstrates a number of known AS genetic associations both within and outside the MHC (*ERAP1*, *IL23R*, chromosome 2p15, chromosome 21q22, *KIF21B-GPR25*, *GPR35*, *IL6R*, *ANTXR2*, *IL1R1*, *FCGR2A* and *NOS2*). It also describes a common novel AS-associated non-MHC variant that achieved genome-wide significance (*CDKAL1*), and three novel common variants that achieved a suggestive level of significance (*FAM118A*, *C7orf72*, *FAM114A1*). One new suggestive rare variant association in *PNPLA1* was identified in single marker analyses, in addition to low-frequency associations observed with rs11465804 in *IL23R* and rs4349859 in *HLA-B\*27*. However, no rare variant associations were noted using burden tests.

In the current study, we identify three AS-associated variants that have previously been associated with IBD, namely variants in

or near *CDKAL1*, *C7orf72* and *TLR10*. There is a high prevalence of IBD in patients with AS;<sup>27,28</sup> around 10% of AS patients have clinical IBD and up to 70% have subclinical bowel inflammation demonstrated histologically.<sup>29</sup> In addition, reactive arthritis, which is a type of spondyloarthritis that can progress to AS, can be triggered by enteric infections such as *Campylobacter*, *Salmonella* and *Shigella*. Strong co-familiality between AS and IBD exists, the sibling recurrence risk ratio for IBD in first-degree relatives of AS probands being 3.0,<sup>30</sup> not dissimilar to the overall familiarity of rheumatoid arthritis.<sup>31</sup> There is extremely strong correlation between AS and IBD genetic associations, with a 2013 analysis indicating that the two diseases shared at that point 22 SNP associations, of which 21 were concordant (same SNP, same direction of association).<sup>32,33</sup> Nonetheless, major differences exist between genetic associations of the two diseases pointing to differences in disease-specific aetiopathogenesis; for example, the absence of association of IBD with *HLA-B\*27*, and the absence of association of AS with the major IBD loci *NOD2/CARD15* and *ATG16L1*. The finding of three more concordant genetic associations further strengthens the evidence of shared aetiopathogenesis between these diseases.

Little is known about *FAM114A1*. It may be responsible for the described association, but the corresponding protein NOXP20 contains a predicted caspase recruitment (CARD) domain suggesting that it may be involved in apoptosis.<sup>34</sup> *CARD9* has previously been associated with AS and NOXP20 may have a similar role. In addition, neighbouring genes include the Toll-like receptors *TLR1* and *TLR6* that are intimately involved in innate immunity and so are strong biological candidates for involvement in AS. The second independent signal at the *FAM114A1* locus was in the toll-like receptor 10 gene (*TLR10*). This missense variant (rs11096955) is predicted by Polyphen-2 to cause a benign change (isoleucine to leucine), but it may tag other more functional variants. This association implicates this important component of the innate immune system in AS aetiology. Toll-like receptors recognise evolutionary conserved sequences on pathogens and trigger immune responses. *TLR10* has been recently identified to induce pro-inflammatory cytokine production and interferon in response to influenza infection.<sup>35</sup> It has recently been suggested that immunodeficiency to gut organisms may trigger AS; if the

association with *TLR10* impairs innate immune responses, this would be consistent with this theory.<sup>36</sup>

*FAM118A* is a protein-coding gene of unknown function that encodes a single-pass transmembrane protein ([www.uniprot.org](http://www.uniprot.org)). The AS-associated SNP (rs6007594) is a missense mutation causing an arginine to be replaced by a histidine. This missense change is predicted to be probably damaging with a Polyphen-2 score of 0.999 (sensitivity: 0.14; specificity: 0.99).<sup>37</sup> Kwan *et al.*<sup>22</sup> demonstrated that *FAM118A* is expressed in lymphoblastoid cell lines as well as human osteoblasts, and in both cell types showed major SNP effects on *FAM118A* expression levels. How this impacts on AS aetiology is not immediately evident and larger studies studying SNPs across this locus will be required to determine the genetic variant(s) responsible for the association observed here.

The *PNPLA1* gene belongs to a family of genes, the members of which have diverse lipolytic and acyltransferase activities. The function of *PNPLA1* itself is not well understood. It is expressed in epidermal keratinocytes, and has a role in glycerophospholipid metabolism in the cutaneous barrier. Variants in *PNPLA1* are associated with the skin disorder ichthyosis.<sup>38</sup> The rs141744967 variant is a missense polymorphism that causes a change from alanine to valine. The functional effect is not available from the Polyphen-2 server; both amino acids are non-polar but differ in size by 28 Da. Further larger studies will be required to determine whether this gene is definitively associated with AS.

This study has several potential limitations; the major limitation is power. The power to detect rare variants is a function of their allele frequency and their effect size along with the population frequency of disease and the required statistical significance threshold. While the study had excellent power to detect common variant associations, the power to detect rare variants was low. Considering variants with a frequency of 0.01 (1%), population disease frequency of 0.005 (0.5%) at a significance threshold of  $P = 5 \times 10^{-8}$  using the participant numbers in this study, the study only had 80% power to detect variants with an additive relative risk of  $> 1.8$ ; other than human leukocyte antigen (HLA) associations, few such variants have been reported in common diseases. This demonstrates poor power to detect individual associations, and increasing the number of cases would improve this power. Nonetheless, if there are large numbers of rare variant associations contributing to AS, the study should have had good power to detect some of these, assuming its coverage of rare variants was good.

The coverage of the Exomechip microarray of rare variants is far from comprehensive, and this impacts both on the coverage of the study, and its ability to pinpoint genetic associations. Further fine-mapping and functional studies will be required to confirm whether the genes we have implicated at each locus are themselves directly involved in AS, or if the SNP associations observed operate by influences on other genes. Sequencing of whole genomes has demonstrated millions of low frequency and rare variants that are not covered on the chip, for example, in the low-coverage analysis of 1000 genomes 15.5 million variants were identified.<sup>12</sup> This suggests that comprehensive rare variant microarray studies may not be feasible, although improvements in imputation methods raise the possibility that many rare but not unique variants may be addressable by this approach.<sup>39</sup>

Finally, rare variants do not share linkage disequilibrium with many other surrounding variants to the extent that common variants do. Therefore if identified in a study such as this, a good check of association is manual inspection of the genotype intensity clustering, and, in addition, considering the biological plausibility. However, probes can map to other areas of the genome without our knowledge giving well-clustered intensity plots. Biological plausibility is not necessarily a good measure of a true association as the association may be the first association in a pathway not previously known to be involved in disease aetiology. This makes independent replication studies essential, although it

is particularly challenging for low-frequency or rare variants because of the sample size requirements.

This study has re-demonstrated many known AS risk loci, and also identified a novel common variant at a genome-wide level of significance, and four suggestive associations, including one rare variant association. The finding of further concordant associations with *IBD* genes increases the evidence of shared aetiopathogenesis between the diseases and the potentially important role of intestinal dysbiosis in AS.<sup>40,41</sup> The major overlap between AS and IBD is also supported by another study showing similar genetic variants but differing effect sizes between variants associated with AS and anterior uveitis.<sup>42</sup> Whether low-frequency and rare genetic variants are major contributors to the aetiopathogenesis of AS remains unclear and will likely require much larger studies with more comprehensive coverage of these variant types to resolve.

## MATERIALS AND METHODS

### Patient cohorts

AS patients of European descent who met the modified New York criteria<sup>43</sup> from the United Kingdom, Australia and New Zealand were recruited ( $n = 5,040$ ). Healthy controls were provided by the following groups (1) 1958 British Birth Cohort ( $n = 5,964$ ); (2) GoDarts type 2 diabetes cohort ( $n = 1,793$ ); (3) Oxford Biobank ( $n = 4,522$ ); (4) Twins UK cohort ( $n = 1,189$ ); (5) Anglo-Australasian Osteoporosis Genetics Consortium ( $n = 7,665$ ). All patients gave written informed consent and ethical approval was provided by all appropriate institutional review boards.

### Genotyping and quality control

Each cohort was genotyped using the Illumina Infinium HumanExome BeadChip version 1.2. This Illumina microarray has ~240,000 markers, made up of exonic variants, splice variants, stop altering variants, ancestry informative markers and MHC tag SNPs. Genotype calling was completed with zCall.<sup>44</sup> Each cohort had QC completed separately, assessing missingness by individual (threshold  $< 3\%$ ), missingness by genotype (threshold  $< 3\%$ ), Hardy-Weinberg equilibrium in controls (Chi-square test threshold  $P = 0.01$ ), extreme heterozygosity (threshold  $> 3$  standard deviations from mean) and identity by descent threshold of PI\_HAT 0.20 was used. After laboratory QC that excluded ~3,000–5,000 SNPs per cohort, 20,714–22,864 SNPs were removed from each set to form a common SNP basis, 1,033 SNPs were removed due to excess missingness, 526 SNPs were removed from control sets due to not being in Hardy-Weinberg equilibrium and 11,711 SNPs were removed due to allele and frequency inconsistencies between the cohorts. 979 subjects were removed due to excess relatedness, 207 were excluded due to extreme heterozygosity and 2 were excluded due to excessive missingness (after SNP QC).

Shared genotyped SNPs between cohorts with  $MAF > 0.05$  were then used to perform principal component analysis for ethnicity identification using SHELLFISH (<http://www.stats.ox.ac.uk/~davison/software/shellfish/shellfish.php>). Unsupervised model-based clustering implemented in R with MCLUST was used to exclude patients deemed to be non-European after plotting with HapMap controls. This model assigns a cluster to each individual based on their principal component analysis values and in consideration to the weighted centre of each cluster and therefore assigns non membership status to those who don't cluster with core HapMap groups. This analysis identified 219 subjects who were removed due to non-European ethnicity. Supplementary Figures 24 and 25 show the principle component analysis after quality control both with and without the addition of Hapmap Samples.

The genomic control measure (GC or  $\lambda$ ) was calculated using the designer suggested random set of 5,000 synonymous variants ([http://genome.sph.umich.edu/wiki/Exome\\_Chip\\_Design#Random\\_set\\_of\\_synonymous\\_variants\\_28as\\_comparator.29](http://genome.sph.umich.edu/wiki/Exome_Chip_Design#Random_set_of_synonymous_variants_28as_comparator.29)).

### Association analysis

For single variant and low-frequency variant analysis, we followed the procedure used by Peloso *et al.*<sup>45</sup> For single-variant analysis, we restricted analysis to variants where the frequency was  $> 0.08\%$ , meaning that 20 or more copies of the minor allele were present. We used Plink to perform association analyses with one eigenvector as a covariate for population

stratification control. Significance levels used were, genome-wide  $P < 5 \times 10^{-8}$ , and suggestive  $5 \times 10^{-8} > P < 5 \times 10^{-6}$ .

For low-frequency variant analysis we used the sequence kernel association test–Optimal (SKAT-O) test that computes the SKAT test<sup>46</sup> and a burden test<sup>47–50</sup> and then selects the test with the best power.<sup>51</sup> This was implemented with the ‘skatMeta’ R package. For the low-frequency variant analysis, we used variants with a frequency of  $< 5\%$ . We also only included gene-based tests where there were at least two variants contributing each with MAF  $> 0.08\%$ , thus ensuring there were at least 40 copies of the minor alleles in the test. Three sets of variants were sequentially used as inputs into the SKAT-O test. This is because non-damaging variants can reduce the power to detect associations in burden tests. The sets used were (a) All variants, (b) ‘Polyphen2 set’: Polyphen-2<sup>37</sup> predicted possibly damaging or probably damaging, and (c) ‘Damaging Set’: Variants causing the following consequences: frameshift substitution, nonframeshift substitution, nonframeshift deletion, nonframeshift insertion, frameshift deletion, frameshift insertion, nonsynonymous single-nucleotide variant, stop-gain single-nucleotide variant, stoploss single-nucleotide variant, missense variant, splice acceptor variant, splice donor variant, splice region variant, initiator codon variant, stop retained variant and incomplete terminal codon variant.

Cluster plots for reported SNPs were checked manually in the case cohort and the 1958 British Birth Cohort. Association of classical alleles was completed with imputed SNP2HLA classical alleles and one principal component by logistic regression in R. The level of significance for this analysis was  $1.2 \times 10^{-4}$ , reflecting Bonferroni correction for the 424 HLA alleles that were tested for association. Conditional analyses were completed by adding the allele being conditioned as a covariate to the logistic regression model. Power calculations were performed with the online Genetic Power Calculator.

## LIST OF WEBSITES

Genetic Power Calculator: (<http://pngu.mgh.harvard.edu/~purcell/gpc/>)

SNAP browser: (<http://www.broadinstitute.org/mpg/snap/>)

## ACKNOWLEDGEMENTS

We acknowledge all participants who gave their DNA for the study. PCR was funded by the National Health and Medical Research Council (NHMRC) of Australia and the RACP-ARA-Starr fellowship. MAB is a Senior Principal Research Fellow of the NHMRC. We acknowledge the Wellcome Trust Case Control Consortium for providing funding for this project. We acknowledge additional financial support from Arthritis Research UK (Grants19536 and 18797) and the NIHR Oxford Comprehensive Biomedical Research Centre Immunity and Inflammation theme (A93081) and the National Ankylosing Spondylitis Society UK.

## COMPETING INTERESTS

The authors declare no conflict of interest.

## REFERENCES

- Robinson, P. C. & Brown, M. A. Genetics of ankylosing spondylitis. *Mol. Immunol.* **57**, 2–11 (2014).
- Baeten, D. *et al.* Anti-interleukin-17A monoclonal antibody secukinumab in treatment of ankylosing spondylitis: a randomised, double-blind, placebo-controlled trial. *Lancet* **382**, 1705–1713 (2013).
- Burton, P. R. *et al.* Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.* **39**, 1329–1337 (2007).
- Evans, D. M. *et al.* Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat. Genet.* **43**, 761–767 (2011).
- Reveille, J. D. *et al.* Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nat. Genet.* **42**, 123–127 (2010).
- Caffrey, M. F. & James, D. C. Human lymphocyte antigen association in ankylosing spondylitis. *Nature* **242**, 121 (1973).
- Brewerton, D. A. *et al.* Ankylosing spondylitis and HL-A 27. *Lancet* **1**, 904–907 (1973).
- Danoy, P. *et al.* Association of variants at 1q32 and STAT3 with ankylosing spondylitis suggests genetic overlap with Crohn’s disease. *PLoS Genet.* **6**, e1001195 (2010).
- Schlossstein, L., Terasaki, P. I., Bluestone, R. & Pearson, C. M. High association of an HL-A antigen, W27, with ankylosing spondylitis. *N. Engl. J. Med.* **288**, 704–706 (1973).
- Robinson, P. C. *et al.* ERAP2 is associated with ankylosing spondylitis in HLA-B27-positive and HLA-B27-negative patients. *Ann. Rheum. Dis.* **74**, 1627–1629 (2015).
- Cortes, A. *et al.* Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat. Genet.* **45**, 730–738 (2013).
- Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
- McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
- Tennesen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
- Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
- Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
- Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn’s disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
- Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn’s disease. *Nat. Genet.* **40**, 955–962 (2008).
- Wolf, N. *et al.* Psoriasis is associated with pleiotropic susceptibility loci identified in type II diabetes and Crohn disease. *J. Med. Genet.* **45**, 114–116 (2008).
- Anderson, C. A. *et al.* Investigation of Crohn’s disease risk loci in ulcerative colitis further defines their molecular relationship. *Gastroenterology* **136**, 523–529, e3 (2009).
- Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
- Kwan, T. *et al.* Tissue effect on genetic control of transcript isoform variation. *PLoS Genet.* **5**, e1000608 (2009).
- Di Meglio, P. *et al.* The IL23R R381Q gene variant protects against immune-mediated diseases by impairing IL-23-induced Th17 effector response in humans. *PLoS ONE* **6**, e17160 (2011).
- Sarin, R., Wu, X. & Abraham, C. Inflammatory disease protective R381Q IL23 receptor polymorphism results in decreased primary CD4+ and CD8+ human T-cell functional responses. *Proc. Natl Acad. Sci. USA* **108**, 9560–9565 (2011).
- Morgan, A. R., Lam, W. J., Han, D. Y., Fraser, A. G. & Ferguson, L. R. Genetic variation within TLR10 is associated with Crohn’s disease in a New Zealand population. *Hum. Immunol.* **73**, 416–420 (2012).
- Abad, C. *et al.* Association of Toll-like receptor 10 and susceptibility to Crohn’s disease independent of NOD2. *Genes Immun.* **12**, 635–642 (2011).
- De Keyser, F. *et al.* Bowel inflammation and the spondyloarthropathies. *Rheum. Dis. Clin. North Am.* **24**, 785–813, ix-x (1998).
- De Vos, M. *et al.* Ileocolonoscopy in seronegative spondylarthropathy. *Gastroenterology* **96**, 339–344 (1989).
- Mielants, H., Veys, E. M., Cuvelier, C., De Vos, M. & Botelberghe, L. HLA-B27 related arthritis and bowel inflammation. Part 2. Ileocolonoscopy and bowel histology in patients with HLA-B27 related arthritis. *J. Rheumatol.* **12**, 294–298 (1985).
- Thjodleifsson, B., Geirsson, A. J., Bjornsson, S. & Bjarnason, I. A common genetic background for inflammatory bowel disease and ankylosing spondylitis: a genealogical study in Iceland. *Arthritis Rheumatol.* **56**, 2633–2639 (2007).
- Frisell, T. *et al.* Familial risks and heritability of rheumatoid arthritis: role of rheumatoid factor/anti-citrullinated protein antibody status, number and type of affected relatives, sex, and age. *Arthritis Rheumatol.* **65**, 2773–2782 (2013).
- Parkes, M., Cortes, A., van Heel, D. A. & Brown, M. A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev. Genet.* **14**, 661–673 (2013).
- Robinson, P. C. & Brown, M. A. The genetics of ankylosing spondylitis and axial spondyloarthritis. *Rheum. Dis. Clin. North Am.* **38**, 539–553 (2012).
- Boucquey, M. *et al.* Noxp20 and Noxp70, two new markers of early neuronal differentiation, detected in teratocarcinoma-derived neuroectodermic precursor cells. *J. Neurochem.* **99**, 657–669 (2006).
- Lee, S. M. *et al.* Toll-like receptor 10 is involved in induction of innate immune responses to influenza virus infection. *Proc. Natl Acad. Sci. USA* **111**, 3793–3798 (2014).
- Kenna, T. J. & Brown, M. Immunopathogenesis of ankylosing spondylitis. *Int. J. Clin. Rheumatol.* **8**, 265–274 (2013).
- Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
- Grall, A. *et al.* PNPLA1 mutations cause autosomal recessive congenital ichthyosis in golden retriever dogs and humans. *Nat. Genet.* **44**, 140–147 (2012).
- Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).

40. Costello, M. E. et al. Intestinal dysbiosis in ankylosing spondylitis. *Arthritis Rheumatol.* **67**, 686–691 (2015).
41. Costello, M.-E., Elewaut, D., Kenna, T. & Brown, M. A. Microbes, the Gut and Ankylosing Spondylitis. *Arthritis Res. Ther.* **15**, 214 (2013).
42. Robinson, P. C. et al. Genetic dissection of acute anterior uveitis reveals similarities and differences in associations observed with ankylosing spondylitis. *Arthritis Rheumatol.* **67**, 140–151 (2015).
43. van der Linden, S., Valkenburg, H. A. & Cats, A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. *Arthritis Rheumatol.* **27**, 361–368 (1984).
44. Goldstein, J. I. et al. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* **28**, 2543–2545 (2012).
45. Peloso, G. M. et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* **94**, 223–232 (2014).
46. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
47. Price, A. L. et al. Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).
48. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**, e1000384 (2009).
49. Morris, A. P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **34**, 188–193 (2010).
50. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
51. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

Supplementary Information accompanies the paper on the *npj Genomic Medicine* website (<http://www.nature.com/npjgenmed>)

### Membership of Wellcome Trust Case Control Consortium 2 (WTCCC2)

#### Management Committee

Peter Donnelly (Chair)<sup>1,2</sup>, Ines Barroso (Deputy Chair)<sup>3</sup>, Jenefer M Blackwell<sup>4,5</sup>, Elvira Bramon<sup>6</sup>, Matthew A Brown<sup>7</sup>, Juan P Casas<sup>8</sup>, Aiden Corvin<sup>9</sup>, Panos Deloukas<sup>3</sup>, Audrey Duncanson<sup>10</sup>, Janusz Jankowski<sup>11</sup>, Hugh S Markus<sup>12</sup>, Christopher G Mathew<sup>13</sup>, Colin NA Palmer<sup>14</sup>, Robert Plomin<sup>15</sup>, Anna Rautanen<sup>1</sup>, Stephen J Sawcer<sup>16</sup>, Richard C Trembath<sup>13</sup>, Ananth C Viswanathan<sup>17</sup>, Nicholas W Wood<sup>18</sup>

#### Data and Analysis Group

Chris CA Spencer<sup>1</sup>, Gavin Band<sup>1</sup>, Céline Bellenguez<sup>1</sup>, Colin Freeman<sup>1</sup>, Garrett Hellenthal<sup>1</sup>, Eleni Giannoulatou<sup>1</sup>, Matti Pirinen<sup>1</sup>, Richard Pearson<sup>1</sup>, Amy Strange<sup>1</sup>, Zhan Su<sup>1</sup>, Damjan Vukcevic<sup>1</sup>, Peter Donnelly<sup>1,2</sup>

#### DNA, Genotyping, Data QC and Informatics Group

Cordelia Langford<sup>3</sup>, Sarah E Hunt<sup>3</sup>, Sarah Edkins<sup>3</sup>, Rhian Gwilliam<sup>3</sup>, Hannah Blackburn<sup>3</sup>, Suzannah J Bumpstead<sup>3</sup>, Serge Dronov<sup>3</sup>, Matthew Gillman<sup>3</sup>, Emma Gray<sup>3</sup>, Naomi Hammond<sup>3</sup>, Alagurevathi Jayakumar<sup>3</sup>, Owen T McCann<sup>3</sup>, Jennifer Liddle<sup>3</sup>, Simon C Potter<sup>3</sup>, Radhi Ravindrarajah<sup>3</sup>, Michelle Ricketts<sup>3</sup>, Matthew Waller<sup>3</sup>, Paul Weston<sup>3</sup>, Sara Widaa<sup>3</sup>, Pamela Whittaker<sup>3</sup>, Ines Barroso<sup>3</sup>, Panos Deloukas<sup>3</sup>.

#### Publications Committee

Christopher G Mathew (Chair)<sup>13</sup>, Jenefer M Blackwell<sup>4,5</sup>, Matthew A Brown<sup>7</sup>, Aiden Corvin<sup>9</sup>, Chris CA Spencer<sup>1</sup>

<sup>1</sup>Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK; <sup>2</sup>Dept Statistics, University of Oxford, Oxford OX1 3TG, UK; <sup>3</sup>Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; <sup>4</sup>Telethon Institute for Child Health Research, Centre for Child Health Research, University of Western Australia, 100 Roberts Road, Subiaco, Western Australia 6008, Australia; <sup>5</sup>Cambridge Institute for Medical Research, University of Cambridge School of Clinical Medicine, Cambridge CB2 0XY, UK; <sup>6</sup>Department of Psychosis Studies, NIHR Biomedical Research Centre for Mental Health at the Institute of Psychiatry, King's College London and The South London and Maudsley NHS Foundation Trust, Denmark Hill, London SE5 8AF, UK; <sup>7</sup>University of Queensland Diamantina Institute, Brisbane, Queensland, Australia; <sup>8</sup>Dept Epidemiology and Population Health, London

School of Hygiene and Tropical Medicine, London WC1E 7HT and Dept Epidemiology and Public Health, University College London WC1E 6BT, UK; <sup>9</sup>Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin 2, Eire; <sup>10</sup>Molecular and Physiological Sciences, The Wellcome Trust, London NW1 2BE, UK; <sup>11</sup>Department of Oncology, Old Road Campus, University of Oxford, Oxford OX3 7DQ, UK, Digestive Diseases Centre, Leicester Royal Infirmary, Leicester LE7 7HH, UK and Centre for Digestive Diseases, Queen Mary University of London, London E1 2AD, UK; <sup>12</sup>Clinical Neurosciences, St George's University of London, London, SW17 0RE, UK; <sup>13</sup>King's College London Dept Medical and Molecular Genetics, King's Heath Partners, Guy's Hospital, London SE1 9RT, UK; <sup>14</sup>Biomedical Research Centre, Ninewells Hospital and Medical School, Dundee DD1 9SY, UK; <sup>15</sup>King's College, London Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Denmark Hill, London SE5 8AF, UK; <sup>16</sup>University of Cambridge Dept Clinical Neurosciences, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK; <sup>17</sup>NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London EC1V 2PD, UK; <sup>18</sup>Dept Molecular Neuroscience, Institute of Neurology, Queen Square, London WC1N 3BG, UK.

### Membership of Australasian Osteoporosis Genetics Consortium (AOGC)

Eugene McCloskey<sup>1</sup>, John Eisman<sup>2</sup>, Graeme Jones<sup>3</sup>, Geoff Nicholson<sup>4</sup>, Richard Eastell<sup>5</sup>, Philip Sambrook<sup>6</sup> (deceased), Richard Prince<sup>7</sup>, Elaine Dennison<sup>8</sup>, Ian Reid<sup>9</sup>, John Wark<sup>10</sup>.

<sup>1</sup>Academic Unit of Bone Metabolism, University of Sheffield, UK; <sup>2</sup>Garvan Institute of Medical Research, University of New South Wales, Australia; <sup>3</sup>Menzies Research Institute, University of Tasmania, Australia; <sup>4</sup>Rural Clinical School, University of Queensland, Australia; <sup>5</sup>Academic Unit of Bone Metabolism, Metabolic Bone Centre, University of Sheffield, UK; <sup>6</sup>Kolling Institute, Royal North Shore Hospital, University of Sydney, Australia; <sup>7</sup>School of Medicine and Pharmacology, University of Western Australia, Australia; <sup>8</sup>Medical Research Council Lifecourse Epidemiology Unit, University of Southampton, UK; <sup>9</sup>Department of Medicine, University of Auckland, New Zealand; <sup>10</sup>Department of Medicine, University of Melbourne, Australia.