DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD

High density genetic mapping identifies new susceptibility loci for rheumatoid arthritis

Citation	Eyre, S., J. Bowes, D. Diogo, A. Lee, A. Barton, P. Martin, A. Zhernakova, et al. 2012. "High density genetic mapping identifies new susceptibility loci for rheumatoid arthritis." Nature genetics 44 (12): 1336-1340. doi:10.1038/ng.2462. http://dx.doi.org/10.1038/ng.2462.
Published Version	doi:10.1038/ng.2462
Accessed	April 17, 2018 4:27:25 PM EDT
Citable Link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:11708546
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

(Article begins on next page)

High density genetic mapping identifies new susceptibility loci for rheumatoid arthritis

Steve Eyre[1,2,24], **John Bowes**[1,2,24], **Dorothée Diogo**[3,4,5,24], **Annette Lee**[6], **Anne Barton**[1,2], **Paul Martin**[1,2], **Alexandra Zhernakova**[7,8], **Eli Stahl**[3,4,5], **Sebastien Viatte**[1,2], **Kate McAllister**[1,2], **Christopher I. Amos**[9], **Leonid Padyukov**[10], **Rene E.M. Toes**[7], **Tom W.J. Huizinga**[7], **Cisca Wijmenga**[8], **Gosia Trynka**[3,4,5,8], **Lude Franke**[8], **Harm-Jan Westra**[8], **Lars Alfredsson**[11], **Xinli Hu**[3,4,5,12], **Cynthia Sandor**[3,4,5], **Paul I.W. de Bakker**[3,4,5,13,14], **Sonia Davila**[15], **Chiea Chuen Khor**[15], **Khai Koon Heng**[15], **Robert Andrews**[16], **Sarah Edkins**[16], **Sarah E Hunt**[16], **Cordelia Langford**[16], **Deborah Symmons**[1,2], **Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate**[17], **Wellcome Trust Case Control Consortium**[17], **Pat Concannon**[18], **Suna Onengut-Gumuscu**[18], **Stephen S Rich**[18], **Panos Deloukas**[16], **Miguel A. Gonzalez-Gay**[19], **Luis Rodriguez-Rodriguez**[20], **Lisbeth Ärlsetig**[21,22], **Javier Martin**[23], **Solbritt Rantapää-Dahlqvist**[21,22], **Robert Plenge**[3,4,5,25], **Soumya Raychaudhuri**[1,2,3,4,5,25], **Lars Klareskog**[10,25], **Peter K Gregersen**[6,25], and **Jane Worthington**[*, 1,2,25]

[1]Arthritis Research UK Epidemiology Unit, Centre for Musculoskeletal Research , University of Manchester, Manchester Academic Health Science Centre [2]National Institute for Health Research Manchester Musculoskeletal Biomedical Research Unit, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Sciences Centre. [3]Division of Rheumatology, Immunology, and Allergy Brigham and Women's, Hospital, Harvard Medical School, Boston, Massachusetts, 02115, USA [4]Division of Genetics, Brigham and Women's, Hospital, Harvard Medical School, Boston, Massachusetts, 02115, USA. [5]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, 02142, USA. [6]The Feinstein Institute for Medical Research, North Shore–Long Island Jewish Health System, Manhasset, New York, USA. [7]Department of Rheumatology, Leiden University Medical Centre, Leiden, The Netherlands. [8]Department of Genetics, University Medical Center Groningen and University of Groningen, Groningen, The Netherlands. [9]University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. [10]Rheumatology Unit, Department of Medicine, Karolinska Institutet and Karolinska University Hospital Solna, Stockholm, Sweden. [11]Department of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden [12]Harvard-MIT Division of Health Sciences and Technology, Boston, Massachusetts [13]Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands [14]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands [15]Division of Human Genetics, Genome Institute of Singapore , Singapore. [16]The Wellcome Trust Sanger Institute, Cambridge, UK [18]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA. [19]Department of

Rheumatology, Hospital Marques de Valdecilla, IFIMAV, Santander, Spain. [20]Hospital Clinico San Carlos, Madrid, Spain. [21]Departments of Public Health and Clinical Medicine Umeå University, Umeå, Sweden [22]Rheumatology, Umeå University, Umeå, Sweden [23]Instituto de Parasitología y Biomedicina López-Neyra, IPBLN-CSIC, Avenida del Conocimiento s/n, Granada, 18100, Spain.

Summary

Using the Immunochip custom single nucleotide polymorphism (SNP) array, designed for dense genotyping of 186 genome wide association study (GWAS) confirmed loci we analysed 11,475 rheumatoid arthritis cases of European ancestry and 15,870 controls for 129,464 markers. The data were combined in meta-analysis with GWAS data from additional independent cases (n=2,363) and controls (n=17,872). We identified fourteen novel loci; nine were associated with rheumatoid arthritis overall and 5 specifically in anti-citrillunated peptide antibody positive disease, bringing the number of confirmed European ancestry rheumatoid arthritis loci to 46. We refined the peak of association to a single gene for 19 loci, identified secondary independent effects at six loci and association to low frequency variants (minor allele frequency <0.05) at 4 loci. Bioinformatic analysis of the data generated strong hypotheses for the causal SNP at seven loci. This study illustrates the advantages of dense SNP mapping analysis to inform subsequent functional investigations.

Rheumatoid arthritis is a common, complex disease affecting up to 1% of the adult population. It is an archetypal autoimmune disease, typified by the presence of serum autoantibodies, including antibodies directed against the Fc portion of immunoglobulins (rheumatoid factor) and against citrullinated peptides (anti-citrillunated peptide antibodies (ACPA)). Genetic studies of rheumatoid arthritis, including recent application of genome wide association studies (GWAS), have identified 32 risk loci among individuals of European ancestry, including *HLA-DRB1*, *PTPN22*, and other loci with shared autoimmune associations[1, 2].

The Immunochip Consortium was formed to design a custom Illumina Infinium array that leveraged the remarkable genetic overlap of susceptibility loci identified across a range of autoimmune diseases. The custom array allows investigators to perform gene-finding and fine-mapping experiments in a co-ordinated manner. Full details have been described previously[3]. Briefly, the array consisted of all known single nucleotide polymorphisms (SNPs) from the 1000 Genomes Project as well as private resequencing efforts for 186 loci, known to be involved in 12 autoimmune diseases. For these loci there is the unique opportunity to fine map autoimmune disease associations. Additional SNPs were included as part of a deep replication effort. This not only provided the opportunity to identify novel rheumatoid arthritis associations with other autoimmune disease loci or with variants with suggestive statistical evidence for association from a previous meta-analysis of but also to refine the GWAS signal and reduce the number of potential causal variants in the 31 non-HLA confirmed loci.

We tested 129,464 polymorphic markers passing quality control, with a minor allele frequency >1%, in 11,475 cases (7,222 ACPA positive, 3,297 ACPA negative and 957 unassigned) and 15,870 controls (Table 1 and Supplementary Tables 1 and 2). We performed analysis on the total rheumatoid arthritis dataset, and also in subsets stratified by ACPA status (Supplementary Table 3). We also had access to GWAS data for an additional 2,363 ACPA positive cases and 17,872 controls, independent of the current study (Table 1). We observed strong evidence of association for the previously identified susceptibility loci (Table 2 and Supplementary Tables 3 and 4).

We identified fourteen novel rheumatoid arthritis loci for populations of European ancestry (*TYK2, IRAK1, TLE3, RASGRP1, PADI4, IL6R, IRF8, ARID5B, IKZF3, RUNX1, POU3F1, RCAN1, CD5, GATA3*) at genome-wide levels of significance ($p<5\times10^{-8}$)(Figure 1): 7 with Immunochip data alone (Table 2) and a further 7 when Immunochip data was combined with the GWAS meta analysis data (Table 2). These loci add 4% to the estimate of heritability explained by confirmed loci, bringing the total to 51%, of which HLA explains 36%. When we removed all known loci from the Immunochip data, we still observed evidence of an excessive number of nominally associated alleles, consistent with the possibility that there are many additional undiscovered alleles [4](Supplementary Figure 1). Interestingly, if a study-wide significance threshold of 9.0×10^{-7} is applied (calculated based on the number of effective independent tests when accounting for linkage disequilibrium (LD)), significant association is also observed at two additional loci; *ELMO1* (rs75351767 $p_{all}=2.94\times10^{-7}$) and *BACH2* (rs72928038 $P_{all} = 8.23\times10^{-7}$) (Supplementary Table 3). A further 8 loci are implicated at suggestive levels of significance ($p<1\times10^{-5}$) in either the full or ACPA positive sub-group analysis including *PTPN2* (rs62097857 $P_{all}=4.4\times10^{-6}$); *TNIP1* (rs6579837 $P_{pos}=1.7\times10^{-6}$) and *TNFSF4* (rs61828284 $P_{pos}=5.4\times10^{-6}$) (Supplementary Table 3).

Previously, we have fine-mapped MHC associations observed in GWAS data of partially overlapping samples by applying imputation of HLA classical alleles and amino acids[5]. The Immunochip platform includes denser SNP coverage within the MHC region which facilitates more accurate imputation. In a preliminary analysis applying the same imputation and fine-mapping approach to ACPA positive cases and controls typed on Immunochip, we observed the same associations that we reported previously. The most significant polymorphic nucleotide was again rs17878703, mapping to position 11 of the *HLA-DRB1* peptide sequence ($p<10^{-677}$). Testing individual amino acid positions within *HLA-DRB1* revealed the strongest association at position 11 ($p<10^{-745}$); conditioning on the position 11 effect we observed association at position 71 ($p=6\times10^{-60}$); finally conditioning on effects at both positions 11 and 71 we observed significant association at position 74 ($p=7\times10^{-19}$). Adjusting for all *HLA-DRB1* alleles to identify independent effects outside this gene we observed significant associations at *HLA-B* corresponding to the presence of aspartate at position 9 in the peptide sequence ($p=1\times10^{-17}$). Adjusting for all *HLA-DRB1* alleles and Asp-9 in *HLA-B*, we observed associations at *HLA-DPB1* corresponding to the presence of phenylalanine at position 9 in the peptide sequence ($p=1\times10^{-17}$).

While it has been demonstrated that ACPA positive and ACPA negative disease has a different allelic association at the MHC and at *PTPN22*[6], previous studies have not been powered to address this issue definitively in additional non MHC loci. Here we analysed 3,297 ACPA negative cases and identify association at genome wide significance to *ANKRD55* (rs71624119 $p=5.2\times10^{-12}$, OR=0.78) in addition to *HLA* (rs4143332 $p=2.9\times10^{-15}$, OR=1.37) (Supplementary Table 3). Strikingly, ANKRD55 has a similar effect as in ACPA positive disease. Comparing association in ACPA positive and negative subgroups we see that for the 45 non-HLA loci, around half show a significantly larger effect size in ACPA positive disease (comparison of OR $p<0.05$), 5 of these loci having a markedly stronger association with this form of disease (*PTPN22, CCR6, CD40, RASGRP1* and *TAGAP*). Eleven loci show no statistical difference in association to either form of rheumatoid arthritis (Supplementary Table 5).This preliminary analysis indicates that differences in the serological subtype of disease may well be reflected in a difference in genetic pre-disposition potentially providing a basis for stratified medicine.

The majority of the 14 new loci associated with rheumatoid arthritis susceptibility, along with previously confirmed loci, were found to contain proteins strongly linked to immune function using GRAIL analysis (Supplementary Table 6 and Supplementary Figure 2), for

example, *CD5, IRF8* and *TYK2*. We also report novel association with *IRAK1*, previously associated with systemic lupus erythematosus (SLE)[7]. This is the first X chromosome locus association to rheumatoid arthritis, and is of relevance given the female predominance of both diseases (9:1 and 3:1 ratio of females: males in SLE and rheumatoid arthritis, respectively). Interestingly, this locus has been shown to occasionally escape X inactivation in female cells[8]. Three of the novel loci confirmed here for the first time in samples of European ancestry have previously been associated in either samples of East Asian ancestry (*PADI4, ARID5B*) or when using a multiethnic approach (*IKZF3*)[9-11]. The SNPs associated in this study are moderately correlated with those identified in samples of East Asian origin, *PADI4* SNPs rs2240336 and rs766449 r2=0.25, D′=1; *ARID5B* SNPs rs12764378 and rs10821944 r2=0.52, D′=0.86. *PADI* genes are involved in the citrullination of peptides and as such are strong candidates for involvement in disease, given the presence of ACPA auto-antibodies. Although the association at *PADI4* (rs2240336) is greater in ACPA positive disease (OR=0.88 P=6.49×10^{-9}) compared to ACPA negative cases (OR=0.93, P=0.01) our formal test comparing OR did not show a statistically significant difference (P=0.14).

We applied conditional logistic regression to test for secondary effects within each locus. In 6 non-*HLA* loci (13%) (*TNFAIP3, CD28, REL, STAT4, TYK2, RASGRP1*) we observed additional independent association signals (Supplementary Figure 3). In total we observed 51 independent risk alleles in 45 non-HLA rheumatoid arthritis loci. To test the possibility that the two risk alleles tag an untyped SNP, we carried out haplotype analysis of the six loci but found no evidence for haplotype specific effects at any locus (Supplementary Table 7). At only four loci, *REL, CD28, TYK2* and *TNFAIP3*, did we observe associations with low frequency variants (MAF<0.05) (Supplementary Table 8).

Out of the 46 rheumatoid arthritis loci, 39 were densely genotyped by Immunochip. For 12 loci we observed that the most strongly associated SNP was not tightly linked to the previously reported leading SNP at that locus, shifting the association signal (Supplementary Table 9). For the 39 confirmed non-HLA rheumatoid arthritis loci on Immunochip, dense mapping refines the association to a single gene for 19 loci (Supplementary Table 10).

Our analysis also identified 7 non-synonymous SNPs within exonic regions (Table 3), as well as a number showing strong regulatory potential (Supplementary Table 11), that are highly correlated (r^2>0.9) with the lead SNP and which are strong candidates for the aetiological variant. The most associated SNP at the *IL6R* locus (rs8192284) is non-synonymous, shows high correlation with circulating IL6R levels and as well as being associated with a decrease risk of coronary heart disease[12, 13], is in strong LD (r^2=0.97, D′=1) with the SNP recently reported to be associated with asthma (rs4129267)[14]. Interestingly, the risk allele at the asthma associated SNP (OR=1.09, p=2.4×10^{-8}) is protective for rheumatoid arthritis (OR=0.9, p = 1.3×10^{-8}). The *IL6R* ligand, *IL6*, is the target of the biologic drug, tocilizumab, which has been shown to be an effective treatment for rheumatoid arthritis. Abatacept is another biologic drug, with therapeutic efficacy in clinical trials and which targets another rheumatoid arthritis susceptibility gene, *CTLA4*. These examples highlight the potential for targeting genes within risk loci.

Testing for statistical interactions between the 46 lead SNPs in confirmed rheumatoid arthritis loci, revealed preliminary evidence for 6 significant pairwise interactions, after Bonferroni correction (p<5×10^{-4}) (Supplementary Table 12). The *GATA3-PRKCQ* interaction is supported by earlier biological observations[15].

From 38 rheumatoid arthritis associated SNPs or proxies accessed for eQTL analysis, 18 showed an eQTL effect on at least one probe, giving a total of 51 SNP-probe combinations with significant eQTL effect (Supplementary Table 13). From these 18 SNPs, 11 showed an

independent or primary eQTL effect on one or more probes (20 SNP-probe combinations), whereas 7 SNPs were not significant after conditioning of the strongest eQTL signal in the locus.

Using a previously described approach, we assessed whether the 46 independent rheumatoid arthritis associated regions, defined by previously known and novel SNP associations discovered here, harboured genes that were specifically expressed in distinct immune cell-types[16]. We observed in a large expression data set of 223 sorted mouse immune cells[17], that these regions contained genes that were most significantly more specifically expressed in CD4+ effector memory T-cell subsets ($p<10^{-7}$) (Supplementary Figure 4).

Of the diseases sharing susceptibility loci with rheumatoid arthritis, systematic fine mapping has only been published, to date, for celiac disease[3]. Previously the two diseases were found to share 6 confirmed non-HLA loci (*MMEL*, *REL*, *CD28/CTLA4*, *TNFAIP3*, *TAGAP* and *IL2/21*) [2]; Immunochip data now identifies an additional 4 confirmed loci common to both diseases (*DDX6, STAT4, PRKCQ* and *IRAK1*) and a further 4 potential rheumatoid arthritis loci (*BACH2*, $p=8.2\times10^{-7}$, *ELMO1*, $p=2.9\times10^{-7}$, *PTPN2*, $p=4.4\times10^{-6}$, *PVT1*, $p=2\times10^{-5}$) in common with confirmed celiac disease loci (Supplementary Table 14). Of the ten rheumatoid arthritis/celiac disease loci, 4 share the same lead SNP (*CD28*, *IL2_21*, *TNFAIP3* and *IRAK1*) and a fifth (*MMEL1*) shares highly correlated SNPs ($r^2>0.88$) and, for all of these variants, the risk allele is the same for both diseases. For two loci (*PRKCQ* and *DDX6*), the lead SNPs are only moderately correlated ($r^2>0.62$) with the minor allele being protective in both diseases. The effects in *STAT4* appear quite different with 3 independent effects in celiac disease and two different independent associations in rheumatoid arthritis. The strongest association signal for risk of celiac disease at *TAGAP* is with the minor allele of a SNP (rs182429) in moderate LD (r2=0.44) with the rheumatoid arthritis risk SNP (rs629326). Indeed, when considering overlap of rheumatoid arthritis susceptibility loci with other autoimmune diseases, only the *PADI4* and *CCL21* non-MHC loci currently show unique association, suggesting that they may be important in determining that the autoimmune reaction is directed at synovial joints.

In summary, through fine mapping on a custom made array designed to capture variation across a number of loci associated with autoimmune diseases, we have identified 14 novel European ancestry rheumatoid arthritis loci; refined the peak of association to a single gene at 19 loci, identified 7 SNPs which might potentially be functional, found independent effects at 6 loci and detected association with SNPs with low MAF (<0.05) at 4 loci. In one third of cases, imputation of GWAS signals without fine-mapping, would have implicated a different genetic region as being disease causal thus illustrating the importance of dense fine mapping analysis prior to embarking on expensive functional studies.

Methods

Genotyping

All samples were genotyped for the Immunochip custom array in accordance with Illumina protocols at six centres: UK (Sanger Centre, Hinxton, Cambridge, UK and the University of Virginia, USA), US and Spain (Feinstein Institute, New York, USA), Sweden EIRA (The Genome Institute, Singapore), Sweden Umea (Department of Medical Sciences, SNP&SEQ Technology Platform, Uppsala University Hospital, Uppsala, Sweden) and The Netherlands (Department of Genetics, University Medical Centre Groningen).

Genotype calling and quality control

Genotype calling was performed on all samples at The University of Manchester as a single project using the Genotyping Module (v1.8.4) of the GenomeStudio Data Analysis Software

package. Initial genotype clustering was performed using the default Illumina cluster file (Immunochip_Gentrain_June2010.egt) and manifest file Immuno_BeadChip_11419691_B.bpm (NCBI build 36) using the GenTrain2 clustering algorithm. Poor performing samples (call rate < 0.90), labelled duplicates (selection informed by 10th percentile GenCall score (p10 GC)) and samples identified post-genotyping as inappropriate for inclusion were also excluded at this point (Supplementary table 1). Automated reclustering was performed on all remaining samples to calibrate clusters on the study sample set.

Poor quality assays were excluded prior to downstream quality control processes by extensive manual review of clustering performance. A subset of good quality SNPs was identified based on the ranking of quality metrics: cluster separation (<0.4), signal intensity (<1.0), call rate (<0.98) and allele frequency. In addition, SNPs that mapped to the Y chromosome or mitochondria, were non-polymorphic, were duplicates, or zeroed in the default Illumina cluster file were also excluded. This resulted in a dataset of 165,549 good quality SNPs (Supplementary Table 2).

To facilitate the meta-analysis and reduce differential missingness each of the six population datasets were processed as discrete entities. SNPs were excluded from each of the datasets with a call rate < 0.99 (cases or controls), a MAF < 0.01 or if they deviated from HWE (p < 5.7×10^{-7}). Samples were excluded with a call rate < 0.99 or if they were identified as outliers based on autosomal heterozygosity (Supplementary Table 3 and 4). Samples were also excluded if they were considered to be outliers based on ethnicity inferred by principal component analysis (PCA). PCA was performed using EIGENSOFT v4.2 with HapMap phase 2 samples as reference populations on a subset of SNPs with a MAF > 0.05 and filtered to minimise inter-marker LD (excluding the MHC region, 23 regions of high LD and previously confirmed rheumatoid arthritis susceptibility regions) (Supplementary Figure 1). Cryptic relatedness was assessed within each dataset by calculating identity-by-descent (IBD) using PLINK v1.07 using the PCA SNP set. A single sample from any related pair (PI_HAT > 0.1875) was removed from the analysis (informed by call rate). In addition IBD was inferred across all six datasets to exclude cross-dataset related individuals (Supplementary table 5). The genomic control inflation factor (λ_{GC}) was calculated within each Immunochip dataset using SNPs included as deep replication for a study investigating the genetic basis for reading and writing ability (submitted by J.C. Barrett). This set of SNPs was filtered as described for the PCA SNP set, leaving a total of 1,469 SNPs distributed evenly across the genome. The λ_{GC} for the datasets was estimated at; 1.07 (UK), 1.03 (US), 0.97 (SE-E), 0.94 (SE-U), 1.12 (NL), and 1.10 (ES). Using the same SNPs to estimate $\lambda GC1000$, where the factor is scaled to the equivalent of 1000 cases and 1000 controls, in the Immunochip meta-analysis resulted in a rescaled λ of 1.02 (1.23 without rescaling).

All novel findings remained significantly associated when including gender and λ_{GC} as a covariate in the analysis (Supplementary Table 15).

Immunochip meta-analysis

Association statistics were calculated in each dataset using logistic regression under an additive model (SNPs coded 0, 1 or 2 with respect to minor allele dosage) and incorporating the top ten principal components as covariates. Odds ratios and standard errors were combined across the six datasets using inverse-variance meta-analysis assuming a fixed effect.

Independent effects

Initial evidence for secondary effects was assessed at each of the previously known and newly identified loci using a forward stepwise logistic regression. The index SNP at each region was included as a covariate and the association statistics re-calculated for the remaining test SNPs. This process was repeated until no SNPs reached the minimum level of significance. The criteria for declaring an independent effect was defined as: p-value < 5×10^{-4}, not highly correlated with index SNP, the conditioned p-value must not differ substantially from the unconditioned value. We next tested if the two-SNP fitted the risk at the locus significantly better than the one-SNP model using a likelihood ratio test.

The effect estimates for each two-SNP haplotype was calculated by including indicator variables for carriage of haplotypes. The indicator variables were constructed by phasing the genotype data for each region satisfying the above criteria were phased using the SHAPEIT algorithm[18].

GWAS meta-analysis

GWAS case-controls collections were previously described[1].Six collections were included in the present study: BRASS, CANADA, EIRA, NARAC1, NARAC2, WTCCC. After quality control and data filtering, the datasets were imputed using IMPUTE and haplotype-phased HapMap Phase 2 European CEU founders as a reference panel[19].

We used IBS estimates to remove related samples across the Immunochip and GWAS collections, using GWAS genotype data instead of imputed data. In each of the twelve collections, we selected a set of SNPs with missing-genotype rate<0.5%, minor allele frequency>5% and Hardy-Weinberg P_{HWE}>5×10^{-7}. Then, we extracted SNPs that passed these filters and were shared between the 12 collections. After further LD pruning and resolving flipping issues, the data from the 12 collections were merged to calculate the IBS statistics. When related samples were identified (siblings or duplicates), the sample from the GWAS collection was removed to preferentially keep Immunochip data in the subsequent association analyses. Filtering and IBS calculation were performed using PLINK[20]. Two GWAS datasets, EIRA and NARAC1, were excluded because of strong overlap (>90% rheumatoid arthritis cases) with the Immunochip SE-E and US collections, respectively. This resulted in a total sample size of 13,838 rheumatoid arthritis cases and 33,742 controls, distributed in 10 collections (Table 1).

The software SNPTEST v2.2 was used to conduct logistic regression analysis of rheumatoid arthritis case-control status in each GWAS collection, conditioning upon the 5 first eigenvectors from PCA analysis, and after excluding SNPs with low statistical information (info score<0.7) or MAF<1%. We also excluded SNPs that were not represented in the filtered Immunochip data. The λ_{GC} for the individual datasets was estimated at; 1.04 (BRASS), 1.02 (CANADA), 1.04 (NARAC2) and 1.05 (WTCCC). There was a slight inflation in λ_{GC} in these cohorts when using the 1,469 SNPs included on Immunochip to investigate the genetic basis of reading and writing ability; 1.11 (BRASS), 1.15 (CANADA), 1.07 (NARAC2) and 1.05 (WTCCC).

We conducted an inverse-variance weighted meta-analysis to combine the results across the 10 collections. We also computed Cochran's Q statistics and I^2 statistics to assess heterogeneity across collections. Meta-analysis and heterogeneity statistics computation was adapted from the MANTEL program[21].

Serological subtype statistical analysis

Multinomial logistic regression was applied to compute odds ratios (OR), 95% confidence interval and p-values for association between the minor allele at every locus and either ACPA-positive ($OR_{ACPA\text{-}positive}$) or ACPA-negative rheumatoid arthritis ($OR_{ACPA\text{-}negative}$) assuming additivity on the log-odds scale (i.e. every locus was coded as 0,1 or 2 corresponding to the copy number of the minor allele). The minor allele was defined according to the allele frequency in the total population, including cases and controls. To test for differences between OR_{ACPA}-positive and OR_{ACPA}-negative, the linear combination $\beta_+ - \beta_-$, where β_+ is log ($OR_{ACPA\text{-}positive}$) and β_- is log ($OR_{ACPA\text{-}negative}$) was calculated, along with its standard error. This enables a p-value for the difference in association to be calculated.

GRAIL analysis

We performed GRAIL analysis (http://www.broadinstitute.org/mpg/grail/grail.php) using HG18 and Dec2006 PubMed datasets, default settings and the 46 genome-wide significant rheumatoid arthritis susceptibility loci (most associated SNP) as seeds.

Interaction analysis

We performed an analysis of epistasis using the most significantly associated SNP from each of the 46 loci (Table 2). Logistic regression was performed in PLINK to model epistasis in each of the the six datasets with the top 10 PCs included as covariates. For each pair of SNPs, the likelihood ratio test was employed to compute the p-value of the interaction term for each dataset. Epistasis results were combined using METAL and Bonferroni corrected.

eQTL analysis

eQTL analysis was done on the peripheral blood of 1,469 unrelated individuals (1,240 samples run on the Illumina HT12v3 platform, 229 samples run on the Illumina H8v2 platform) from the United Kingdom and the Netherlands. Details of the eQTL analysis have been previously described[22] . In short, we assessed the effect of all rheumatoid arthritis associated SNPs (Table 2) on expression of genes, located within 250kb left and right from the SNP (cis eQTLs).

All individuals from the eQTL study were genotyped on Illumina Hap300K platform and then imputed to HapMap 2 using Impute 2.0 software. Since not all SNPs from Illumina Immunochip platform were genotyped or imputed on the 1,469 eQTL samples, we used the following strategy (Supplementary Figure 5): First, we investigated whether the SNP is present in the eQTL data and had passed the QC for eQTL mapping (MAF >= 5%, HWE P-value >= 0.001, call rate >= 95%). From 50 rheumatoid arthritis-SNPs, 26 were present in HapMap imputed datasets and were directly assessed for eQTL effects (Supplementary Table 13). For the other 24 SNPs, not present in our HapMap imputed data, we checked whether the rheumatoid arthritis-SNP was available in 1000 genomes database. If so, we queried all SNPs within 10MB of the rheumatoid arthritis-SNP that were also present in the eQTL data and would pass eQTL QC measures, and picked the SNP with the highest LD present in HapMap after QC. The threshold of $r^2 > 0.8$ for the LD was used. For 12 SNPs, no proxy was available with our criteria, and these SNPs were not included in the eQTL analysis. For the remaining 12 SNPs the best proxy SNP is included to the eQTL table (Supplementary Figure 5).

We also performed a cis-eQTL analysis for the top associated gene expression probe, as well as two conditional analyses: (1) conditioning on the effect of the rheumatoid arthritis-SNP

(gSNP), and (2) conditioning on the effect of the top eQTL SNP (eSNP) (Supplementary Table 13).

The rheumatoid arthritis associated SNP was labelled as having a primary effect on gene expression if it was either the top eQTL in the locus, or was a good proxy of top eQTL SNP (r2>8). It was labelled as an independent eQTL if it showed an effect after conditioning on the primary eQTL. From 20 rheumatoid arthritis SNPs, that showed an eQTL effect, 13 had either an independent or primary eQTL effect on one or more probes (22 SNP-probe combinations). A further 7 SNPs were not significant after the conditioning of the strongest eQTL signal in the locus, suggesting that they are not primary eQTLs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Reference List

1. Stahl EA, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. 2010; 42:508–514. [PubMed: 20453842]

2. Zhernakova A, et al. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS. Genet. 2011; 7:e1002004. [PubMed: 21383967]

3. Trynka G, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat. Genet. 2011; 43:1193–1201. [PubMed: 22057235]

4. Stahl EA, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat. Genet. 2012 [PubMed: 22446960]

5. Raychaudhuri S, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat. Genet. 2012; 44:291–296. [PubMed: 22286218]

6. Padyukov L, et al. A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis. Ann. Rheum. Dis. 2011; 70:259–265. [PubMed: 21156761]

7. Jacob CO, et al. Identification of IRAK1 as a risk gene with critical role in the pathogenesis of systemic lupus erythematosus. Proc. Natl. Acad. Sci. U. S. A. 2009; 106:6256–6261. [PubMed: 19329491]

8. Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature. 2005; 434:400–404. [PubMed: 15772666]

9. Suzuki A, et al. Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat. Genet. 2003; 34:395–402. [PubMed: 12833157]

10. Kurreeman FA, et al. Use of a Multiethnic Approach to Identify Rheumatoid-Arthritis-Susceptibility Loci, 1p36 and 17q12. Am. J. Hum. Genet. 2012; 90:524–532. [PubMed: 22365150]

11. Okada Y, et al. Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. Nat. Genet. 2012 [PubMed: 22446963]

12. Hingorani AD, Casas JP. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. Lancet. 2012; 379:1214–1224. [PubMed: 22421340]

13. Sarwar N, et al. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. Lancet. 2012; 379:1205–1213. [PubMed: 22421339]

14. Ferreira MA, et al. Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. Lancet. 2011; 378:1006–1014. [PubMed: 21907864]

15. Stevens L, et al. Involvement of GATA3 in protein kinase C theta-induced Th2 cytokine expression. Eur. J. Immunol. 2006; 36:3305–3314. [PubMed: 17111354]

16. Hu X, et al. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. Am. J. Hum. Genet. 2011; 89:496–506. [PubMed: 21963258]

17. Heng TS, Painter MW. The Immunological Genome Project: networks of gene expression in immune cells. Nat. Immunol. 2008; 9:1091–1094. [PubMed: 18800157]

18. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat. Methods. 2012; 9:179–181. [PubMed: 22138821]

19. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 2007; 39:906–913. [PubMed: 17572673]

20. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007; 81:559–575. [PubMed: 17701901]

21. de Bakker PI, et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet. 2008; 17:R122–R128. [PubMed: 18852200]

22. Fehrmann RS, et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS. Genet. 2011; 7:e1002197. [PubMed: 21829388]
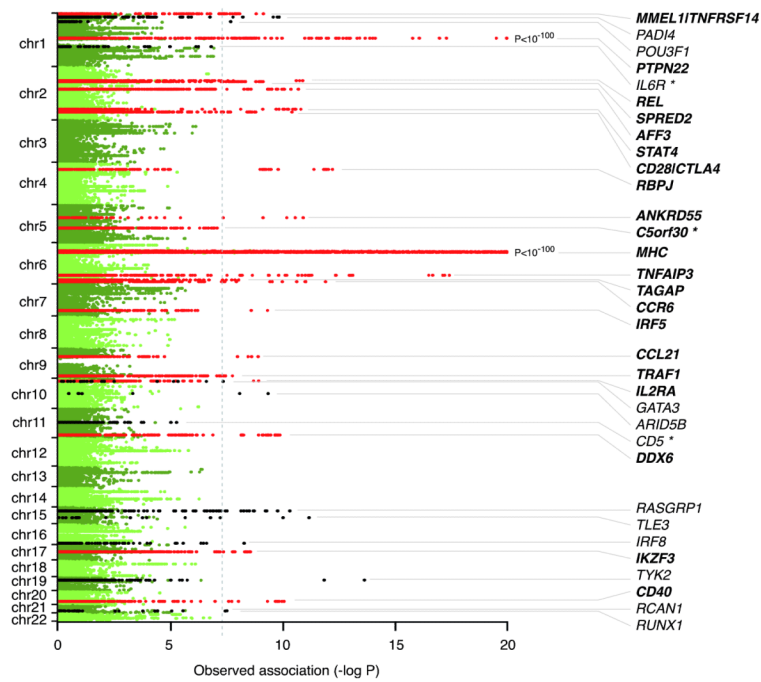
Figure 1. Manhattan plot of association statistics highlighting all autosomal loci associated to rheumatoid arthritis in the study

P values of association to ACPA positive rheumatoid arthritis from the meta-analysis of the Immunochip and GWAS data are shown. Known and new rheumatoid arthritis associated loci are shown in red and black respectively. Three associated loci (identified by a *) only reach $P<5\times10^{-8}$ when ACPA positive and ACPA negative cases are included in the analysis. The dashed grey line indicates genome-wide significance ($P=5\times10^{-8}$).

Table 1

Sample Collections

Rheumatoid arthritis cases and controls for the Immunochip analysis were assembled from a number of different studies from 6 centres across 5 countries (online methods). Genotype data for additional samples analysed in previously published rheumatoid arthritis ACPA positive GWAS were available from 4 studies. Rheumatoid arthritis cases were classified as anti citrullinated peptide antigen (ACPA) positive and ACPA negative. F:M = female: male.

	Collection	Cases				Controls	
		All	% Female	ACPA +	ACPA −		% Female
Immunochip	UK	3870	74	2406	1000	8430	53
	Swedish EIRA	2762	70	1762	987	1940	73
	US	2536	75	1803	593	2134	65
	Dutch	648	66	330	301	2004	42
	Swedish Umea	852	70	524	242	963	69
	Spanish	807	74	397	216	399	65
	TOTAL	11475	73	7222	3339	15870	57
GWAS	BRASS (US)	479	82	479	-	1627	45
	Canada	586	76	586	-	1553	54
	NARAC2 (US)	746	48	746	-	6567	49
	WTCCC (UK)	552	74	552	-	8125	46
	TOTAL	2363	68	2363	-	17872	48
TOTAL		13838	72	9585	3297	33742	52

Table 2
Non-HLA loci associated with rheumatoid arthritis at genome-wide significance level

Novel loci are shown with either the best SNP on Immunochip (Table 2a), if $<5\times10^{-8}$, or from the most associated SNP from the combined analysis of GWAS and Immunochip data (Table 2b).

SNP	Gene	Chr	MAF	Risk allele	P	OR	LD region r²>0.9*	Region size	Localization of LD region (r²>0.9) relative to nearest genes
					Novel loci on Immunochip				
rs34536443[b]	TYK2	19p13	0.04	G	2.3×10^{-14}	0.62	10,427,721-10,492,274	64.55 kb	47.96 kb 5′ to exon 13 of RAVER1; complete ICAM3; complete TYK2
rs13397[b]	IRAK1	Xq28	0.12	A	1.2×10^{-12}	1.27	153,196,345-153,248,248	51.9kp	5′ to exon 2 of TMEM187; HCFC1; 25 kb 3′ of IRAK1
rs8026898[b]	TLE3 #	15q23	0.29	A	1.4×10^{-10}	1.17	69,984,462-70,010,647	26.19 kb	329.48 kb 3′ of TLE3
rs8043085[b]	RASGRP1	15q14	0.25	A	1.4×10^{-10}	1.17	38,828,140-38,844,106	15.97 kb	Intron 2 of RASGRP1;
rs2240336[b]	PADI4	1p36	0.42	A	5.9×10^{-09}	0.88	17,673,102-17,674,402	1.30 kb	Intron 9 PADI4
rs8192284[a] (rs2228145)	IL6R	1q21	0.42	C	1.3×10^{-08}	0.9	154,418,749-154,428,283	9.54 kb	Intron 6 to intron 9 of IL6R
rs13330176[b]	IRF8	16q24	0.22	A	4.0×10^{-08}	1.15	86,016,026-86,019,087	3.06 kb	59.83 kb 3′ of IRF8
					Novel loci adding GWAS data				
rs12764378[d]	ARID5B #	10q21	0.23	A	4.5×10^{-10}	1.14	63,786,554-63,800,004	13.45 kb	Intron 4 of ARID5B
rs9979383[c]	RUNX1 #	21q22	0.36	G	5.0×10^{-10}	0.9	36,712,588-36,715,761	3.17 kb	5′ region of RUNX1
rs12936409/ rs2872507[c]	IKZF3	17q12	0.47	A	2.8×10^{-09}	1.1	37,912,377-38,080,912	168.54 kb	IKZF3;GSDMB; Intron 1 to 164.92 kb 3′ of ORMDL3
rs883220[d]	POU3F1 #	1p34	0.26	A	2.1×10^{-08}	0.89	38,614,867-38,644,861	30.00 kb	102.42 kb 5′ of POU3F1
rs2834512[d]	RCAN1 #	21q22	0.12	A	2.1×10^{-08}	0.86	35,909,625-35,930,915	21.29 kb	Intron 1 of RCAN1
rs595158[c]	CD5	11q12	0.49	C	3.4×10^{-08}	1.09	60,888,001-60,922,634	34.63 kb	Intron 5 to 27.31 kb 3′ of CD5; Intron 1 to 9.73 kb 3′ of VPS37C
rs2275806[d]	GATA3 #	10p14	0.41	G	4.6×10^{-08}	1.11	8,095,340-8,097,368	2.03 kb	227bp 5′ to exon 2 of GATA3
					Known loci on Immunochip				
rs2476601[b]	PTPN22	1p13	0.09	A	7.5×10^{-77}	1.78	114,303,808-114,377,568	73.76 kb	Complete RSBN1; Exon 14 to 52.62 kb 3′ of PTPN22
rs71624119[a]	ANKRD55	5q11	0.25	A	5.6×10^{-20}	0.81	55,440,730-55,442,249	1.52 kb	Intron 6 of ANKRD55
rs6920220[b]	TNFAIP3	6q23	0.2	A	2.3×10^{-13}	1.2	137,959,235-138,006,504	47.27 kb	181.85 kb 5′ of TNFAIP3
rs932036[a]	RBPJ	4p15	0.3	A	2.0×10^{-10}	1.14	26,085,480-26,128,710	43.23 kb	36.37 kb 5′ of RBPJ
rs59466457[b]	CCR6	6q27	0.44	A	2.7×10^{-10}	1.15	167,526,096-167,540,842	14.75 kb	Intron 1 of CCR6
rs13426947[a]	STAT4	2q32	0.19	A	7.2×10^{-10}	1.15	191,900,449-191,935,804	35.36 kb	Intron 5 to 18 of STAT4

SNP	Gene	Chr	MAF	Risk allele	P	OR	LD region $r^2>0.9$*	Region size	Localization of LD region ($r^2>0.9$) relative to nearest genes
rs2812378[b]	*CCL21*	9p13	0.34	G	7.2×10^{-10}	1.15	34,707,373-34,710,338	2.97 kb	CCL21
rs6032662[b]	*CD40*	20q13	0.24	G	1.4×10^{-09}	0.86	44,730,245-44,747,947	17.70 kb	16.67 kb 5′ to intron 1 of CD40
rs2843401[b]	*MMEL1*	1p36	0.33	A	6.6×10^{-09}	0.87	2,516,781-2,709,164	192.38 kb	Complete MMEL1,C1ORF93; TTC34
rs10209110[a]	*AFF3*	2q11	0.49	A	1.1×10^{-08}	0.9	100,640,432-100,730,111	89.68 kb	5′ region to intron 2 of AFF3
rs34695944[b]	*REL*	2p16	0.37	G	2.6×10^{-08}	1.13	61,072,664-61,164,331	91.67 kb	Complete REL
rs11571302[b]	*CTLA4*	2q33	0.48	A	4.5×10^{-08}	0.89	204,738,919-204,745,003	6.08 kb	236bp 3′ of CTLA4; 56.47 kb 5′ of ICOS;
rs39984[a]	*GIN1*	5q21	0.32	A	9.3×10^{-08}	0.88	102,595,778-102,625,335	29.56 kb	Intron 1 to 10.97 kb 3′ of C5orf30; 139.92 kb 5′of GIN1
rs35677470[a]	*DNASE1L3*	3p14	0.08	A	1.7×10^{-07}	1.19	58,181,499-58,183,636	2.14 kb	Exon 8 to intron 9 of DNASE1L3; 134.97 kb 5′ of PXK
rs3807306[b]	*IRF5*	7q32	0.49	C	1.9×10^{-07}	0.89	128,580,680-128,580,680	1bp	Intron 1 of IRF5
rs3218251[b]	*IL2RB*	22q12	0.25	A	1.9×10^{-07}	1.13	37,544,245-37,545,505	1.26 kb	Intron 1 of IL2RB
rs4938573[b]	*DDX6*	11q23	0.18	G	5.3×10^{-07}	0.87	118,662,993-118,745,884	82.89 kb	Complete SETP16; 1.14 kb 5′ of DDX6
rs6546146[b]	*SPRED2*	2p14	0.38	A	8.0×10^{-07}	0.9	65,556,324-65,598,300	41.98 kb	Intron 1 to intron 4 of SPRED2
rs629326[b]	*TAGAP*	6q25	0.41	C	1.1×10^{-06}	0.9	159,489,791-159,496,713	6.92 kb	23.61 kb 5′ of TAGAP
rs10739580[b]	*TRAF1*	9q33	0.33	G	1.7×10^{-06}	1.12	123,640,500-123,708,286	67.79 kb	Complete TRAF1
rs10795791[a]	*IL2RA*	10p15	0.4	G	3.0×10^{-06}	1.09	6,106,266-6,108,340	2.08 kb	1.93 kb 5′ of IL2RA
rs4840565[a]	*BLK*	8p23	0.27	G	3.9×10^{-06}	1.1	11,338,383-11,352,485	14.10 kb	13.13 kb 5′ to intron 1 of BLK
rs798000[b]	*CD2*	1p13	0.34	G	6.2×10^{-06}	1.11	117,280,696-117,280,696	1bp	16.31 kb 5′ of CD2
rs1980422[f]	*CD28*	2q33	0.23	G	8.7×10^{-06}	1.12	204,610,004-204,634,569	24.57 kb	7.45 kb 3′ of CD28; 97.94 kb 5′ of CTLA4
rs2014863[a]	*PTPRC*	1q31	0.36	C	2.1×10^{-05}	1.09	198,791,907-198,810,008	18.10 kb	65.36 kb 3′ of PTPRC
rs10683701[b]	*KIF5A*	12q13	0.33	-	2.3×10^{-05}	0.9	58,034,835-58,105,094	70.26 kb	snoU13;52.90 kb 5′ to intron 5 of OS9; 54.42 kb 3′ of KIF5A
rs947474[a]	*PRKCQ*	10p15	0.17	G	2.5×10^{-05}	0.9	6,390,450-6,390,450	1bp	78.66 kb 3′ of PRKCQ
rs10494360[b]	*FCGR2A*	1q23	0.12	G	3.0×10^{-05}	1.14	161,463,876-161,480,649	16.77 kb	11.34 kb 5′ to exon 5 of FCGR2A
rs6911690[b]	*PRDM1*	6q21	0.12	G	1.2×10^{-04}	0.87	106,435,981-106,508,640	72.66 kb	25.55 kb 5′ of PRDM1
rs78560100[a]	*IL2-IL21*	4q27	0.07	C	5.8×10^{-04}	1.13	123,030,583-123,503,591	473.01 kb	KIAA1109;ADAD1;IL2; 30.19 kb 3′ of IL21
rs570676[b]	*TRAF6*	11p12	0.38	A	2.1×10^{-03}	0.93	36,486,064-36,519,624	33.56 kb	Intron 3 to 22.51 kb 3′ of TRAF6

[a]Previously indentified loci are shown with the most significantly associated SNP on Immunochip (2c) indicates the data is from all rheumatoid arthritis samples on Immunochip

[b]Previously indentified loci are shown with the most significantly associated SNP on Immunochip (2c) is data from Immunochip for ACPA positive individuals

[c]Previously indentified loci are shown with the most significantly associated SNP on Immunochip (2c) is data from adding GWAS samples and all rheumatoid arthritis Immunochip data

[d]Previously indentified loci are shown with the most significantly associated SNP on Immunochip (2c) is from ACPA positive Immunochip and GWAS data.

[*] co-ordinates based on GRCh37 assembly.

[#] region not included for dense mapping on Immunochip.

Table 3

Potential causal exonic SNPs located by Immunochip dense genotyping. Conservation is by phastCons17way, study 99th percentile = 0.998; 95th percentile = 0.367. An **essential splice site** is a splice donor variant within the 2 base pair region at the 5′ end of an intron. A **splice site** is a sequence variant within 1-3 base pairs of the exon or 3-8 base pairs of the intron.

Chr	POS	Gene	SNP	MAF	r^2 with lead	SNP Location	Allele	Amino acid change	Polyphen	SIFT	Conservation
1	2,535,613	MMEL1	rs4648562	0.33	1	Essential splice site	A	-	-	-	1
1	114,377,568	PTPN22	rs2476601	0.12	lead	Non-Synonymous coding	A	Arg620Trp	benign	tolerated	0.992
1	154,426,970	IL6R	rs2228145	0.38	lead	Non-Synonymous coding	C	Asp358Ala	benign	tolerated	0.008
3	58,183,636	DNASEIL3	rs35677470	0.09	lead	Non-Synonymous coding	A	Arg206Cys	probably damaging	deleterious	0.992
11	60,893,235	CD5	rs2229177	0.47	0.96	Non-Synonymous coding	C	Ala471Val	probably damaging	deleterious	0.835
19	10,449,358	ICAM3	rs7258015	0.23	1	Non-Synonymous coding/ Splice site	C	Arg115Gly	benign	tolerated	0
19	10,463,118	TYK2	rs34536443	0.04	lead	Non-Synonymous coding	G	Pro1104Ala	probably damaging	deleterious	0.189