ARTHRITIS & RHEUMATISM
Vol. 65, No. 12, December 2013, pp 3026–3035
DOI 10.1002/art.38110

A Genome-Wide Association Study Reveals *ARL15*, a Novel Non-HLA Susceptibility Gene for Rheumatoid Arthritis in North Indians

Sapna Negi,[1] Garima Juyal,[2] Sabyasachi Senapati,[2] Pushplata Prasad,[2] Aditi Gupta,[2] Shalini Singh,[2] Sujit Kashyap,[2] Ashok Kumar,[3] Uma Kumar,[3] Rajiva Gupta,[3] Satbir Kaur,[3] Suraksha Agrawal,[4] Amita Aggarwal,[4] Jurg Ott,[5] Sanjay Jain,[6] Ramesh C. Juyal,[1] and B. K. Thelma[2]

Objective. **Genome-wide association studies (GWAS) and their subsequent meta-analyses have changed the landscape of genetics in rheumatoid arthritis (RA) by uncovering several novel genes. Such studies are heavily weighted by samples from Caucasian populations, but they explain only a small proportion of total heritability. Our previous studies in genetically distinct North Indian RA cohorts have demonstrated apparent allelic/genetic heterogeneity between North Indian and Western populations, warranting GWAS in non-European populations. We undertook this study to detect additional disease-associated loci that may be collectively important in the presence or absence of genes with a major effect.**

Methods. **High-quality genotypes for >600,000 single-nucleotide polymorphisms (SNPs) in 706 RA patients and 761 controls from North India were generated in the discovery stage. Twelve SNPs showing suggestive association ($P < 5 \times 10^{-5}$) were then tested in an independent cohort of 927 RA patients and 1,148 controls. Additional disease-associated loci were determined using support vector machine (SVM) analyses. Fine-mapping of novel loci was performed by using imputation.**

Results. **In addition to the expected association of the HLA locus with RA, we identified association with a novel intronic SNP of *ARL15* (rs255758) on chromosome 5 ($P_{combined} = 6.57 \times 10^{-6}$; odds ratio 1.42). Genotype–phenotype correlation by assaying adiponectin levels demonstrated the functional significance of this novel gene in disease pathogenesis. SVM analysis confirmed this association along with that of a few more replication stage genes.**

Conclusion. **In this first GWAS of RA among North Indians, *ARL15* emerged as a novel genetic risk factor in addition to the classic HLA locus, which suggests that population-specific genetic loci as well as those shared between Asian and European populations contribute to RA etiology. Furthermore, our study reveals the potential of machine learning methods in unraveling gene–gene interactions using GWAS data.**

Supported by grants from the Department of Biotechnology, Government of India through the Centre of Excellence in Genome Sciences and Predictive Medicine (BT/01/COE/07/UDSC/2008) and Methotrexate in Rheumatoid Arthritis: Pharmacogenetics and Clinico-Immunological Correlates (BT/PR5356/Med/14/624/2004) to Drs. A. Kumar, U. Kumar, R. Gupta, Jain, R. C. Juyal, and Thelma, and by grant 30730057 to Dr. Ott from the National Natural Science Foundation of China, Beijing, China. Mr. Senapati and Ms A. Gupta are recipients of senior research fellowships from the Council of Scientific and Industrial Research, New Delhi, India.

[1]Sapna Negi, PhD (current address: Regional Medical Research Centre, Bhubaneswar, India), Ramesh C. Juyal, PhD: National Institute of Immunology, New Delhi, India; [2]Garima Juyal, PhD, Sabyasachi Senapati, MPhil, Pushplata Prasad, PhD, Aditi Gupta, MSc, Shalini Singh, PhD, Sujit Kashyap, MTech, B. K. Thelma, PhD: University of Delhi South Campus, New Delhi, India; [3]Ashok Kumar, MD, FRCP (current address: Fortis Flt. Lt. Rajan Dhall Hospital, New Delhi, India), Uma Kumar, MD, Rajiva Gupta, MD, FRCP (current address: Medanta–The Medicity, Gurgaon, India), Satbir Kaur, MBBS, DNB (General Medicine): All India Institute of Medical Sciences, New Delhi, India; [4]Suraksha Agrawal, PhD, Amita Aggarwal, MD, DM: Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, India; [5]Jurg Ott, PhD: Chinese Academy of Sciences, Beijing, China; [6]Sanjay Jain, PhD: University of Delhi, Delhi, India, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India, and Santa Fe Institute, Santa Fe, New Mexico.

Drs. Negi and G. Juyal contributed equally to this work. Mr. Senapati and Dr. Prasad contributed equally to this work. Ms A. Gupta and Dr. Singh contributed equally to this work.

Address correspondence to B. K. Thelma, PhD, Department of Genetics, University of Delhi South Campus, Benito Juarez Marg, New Delhi 110021, India. E-mail: thelmabk@gmail.com.

Submitted for publication March 13, 2013; accepted in revised form July 25, 2013.

Rheumatoid arthritis (RA) constitutes an important and growing public health burden in both western and developing countries. Cumulative epidemiologic surveys have shown that genetic susceptibility plays an important role (1,2). Over the last few years, a spate of genome-wide association studies (GWAS) and their subsequent meta-analyses, comprising large cohort sizes, have been successful in uncovering several novel disease genes in complex traits, with RA and inflammatory bowel disease (IBD) genetics being the principal beneficiaries (3–6). These studies have undoubtedly provided valuable insights into the genetic architecture of RA by unearthing and deciphering unique novel genes and pathways as well as those shared with other autoimmune disorders. To date, 46 genes/loci for susceptibility to RA have been confirmed, primarily in individuals of European ancestry (4). Only ~4% of all GWAS have been performed in populations of non-European origin, and findings of these GWAS have been disparate from those of GWAS performed in Caucasians (7–9). In this regard, our previous studies in a North Indian RA cohort have demonstrated limited replication of European GWAS findings (10). Furthermore, the 2 non–major histocompatibility complex (non-MHC) genes associated with RA, namely, the *PADI4* locus in Asian populations and *PTPN22* in Europeans, have never had a similar effect in these 2 ethnic groups. The associations with *PADI4* are extremely weak or absent in most European studies. Conversely, the *PTPN22* risk allele is very rare in Asian populations (11).

These observations have raised the question of whether markers are generalizable across populations, and they have emphasized the need to conduct GWAS in populations of distinct ethnicities. Such studies might not only reveal "ethnicity-specific" genes but might also explain a small proportion of the "total heritability," explain the performance of available Caucasian population–based genotyping platforms in ethnically diverse groups, and aid in cross-ethnicity–based fine mapping. Herein we report results from the first GWAS of RA conducted in the genetically distinct North Indian population. We identified 1 novel gene, namely, *ARL15*, in addition to reconfirming a few previously reported GWAS findings.

Meta-analyses of GWAS have extended the success of complex disease genetics by revealing novel variants with smaller effect sizes. Despite this, their findings account for only a small fraction of the total disease heritability. This suggests either that there may be undiscovered genetic mechanisms or that alternative methods are required to analyze the rich genome-wide resource to address the missing heritability. Therefore, in parallel with GWAS, we employed the other approach of support vector machine (SVM) analysis to detect additional disease-associated loci that may be collectively important in the presence or absence of genes with a major effect.

SUBJECTS AND METHODS

Study participants. Two independent case–control sample sets were used in the discovery and replication stages. The discovery set consisted of 706 RA patients and 761 controls, and the replication set consisted of 927 RA patients and 1,148 controls. Controls were age-, sex-, and ethnicity-matched healthy unrelated blood donors with no history of chronic inflammatory autoimmune or infectious diseases. RA patients were predominantly female (further information is available from the corresponding author). All individuals were self-reported North Indians. RA was diagnosed according to the 1987 revised classification criteria of the American College of Rheumatology (12). A more detailed description of subjects is available from the corresponding author. Informed consent was obtained from each participant, and approval for the study was obtained from the ethics committees of appropriate institutions.

Genotyping. Genomic DNA was extracted from peripheral whole blood using a standard phenol–chloroform protocol. Genotyping for the GWAS stage was carried out at Sandor Proteomics with an Illumina CSPro, using an Illumina Human660W Quad BeadChip genotyping platform. Data were imported into GenomeStudio software for initial review and quality control. Cluster statistics were recalculated using 130 North Indian samples from our data set as reference. For the replication stage, genotyping was performed using a MassArray system (Sequenom) by AceProbe Technologies, a commercial facility.

Statistical analysis. Data were analyzed using Plink software (http://pngu.mgh.harvard.edu/~purcell/plink) and R software (http://www.r-project.org). GWAS genotype data were subjected to quality control. Samples with a call rate of <95%, ambiguous sex, ethnic outliers (identified by multidimensional scaling plots), duplicates, and first-degree relatives (pairwise identity by descent score [pi-hat] = 0.4) were excluded. Single-nucleotide polymorphisms (SNPs) lying in X-chromosomal, Y-chromosomal, and mitochondrial regions were excluded from the analysis. Multidimensional scaling was performed using 12,000 population-differentiating markers for identifying population stratification (13). The over-dispersion factor of association test statistics (genomic control inflation factor), λ_{GC}, was calculated using observed versus expected values for all SNPs. We excluded SNPs with a call rate of <95%, a minor allele frequency (MAF) of <5%, or deviation from Hardy-Weinberg equilibrium ($P_{hwe} < 1 \times 10^{-7}$) in the controls using Plink software (14) (details are available from the corresponding author). Allelic association was calculated using the Cochran-Armitage trend test and the additive dominant and recessive test, and the risk of disease was estimated using the odds ratio. *P* values less than 5×10^{-8} and less than 5×10^{-5} were considered significant for disease association

and suggestion of disease association, respectively, at the genome-wide level. Manhattan plots for chromosome-wise distribution of association *P* values were generated using R software.

For the subsequent replication study, we considered SNPs with *P* values less than 5×10^{-5} in the discovery stage in allelic and additive dominant/recessive models. Of these, the previously reported genes and the HLA region were excluded. Furthermore, except at LOC150577, only 1 SNP per gene/locus with the strongest association was selected whenever multiple hits were found. Genotyping was carried out in an independent sample set comprising 927 RA patients and 1,148 controls. As in the genome-wide study, SNPs with a call rate of <95%, an MAF of <5%, or deviation from Hardy-Weinberg equilibrium ($P_{hwe} < 1 \times 10^{-7}$) in the controls were removed from the replication stage.

SVM/multivariate analysis. SVM analysis (15,16) is a machine learning method that uses "training data" consisting of samples of multivariate input with known classification to construct a mathematical "model" or "classifier" that can then be used for classifying a new multivariate input. The method has been used in conjunction with GWAS (17–20), where the training data consist of the set of alleles present at multiple SNP locations in the genome of several known patients and controls. Using these data, SVM analysis constructs a model that can predict the disease status of an individual from his or her genotype. The model is obtained by representing each individual in the training set as a point in an *n*-dimensional "feature space" (*n* is the number of SNPs) and constructing an $n - 1$–dimensional hypersurface in this space that best separates the points corresponding to patients and controls. Since the model uses the genotype at a large number of loci to make the prediction, it implicitly integrates information about the collective effect of multiple loci and thus provides information complementary to the chi-square statistic, which evaluates the significance of a single SNP. Furthermore, the model also assigns a weight to each SNP that is a measure of its importance in predicting the disease status, and thus the model can be used to rank SNPs.

The performance or success rate in correctly classifying new inputs of SVM models depends on the size of the training set (number of individuals) compared to the number of features (number of SNPs). For a reasonable performance, the dimensionality of the feature space should not be too large. We therefore imposed a cutoff *P* value of less than or equal to 0.001 to prune the set of SNPs considered. This yielded a set of 517 SNPs, which is a reasonable number to consider given the size of our data set (1,112 individuals). We constructed 100 different subsets of the data, each consisting of 90% of the individuals chosen randomly from 1,112 individuals, and for each subset we constructed an SVM model using the publicly available package LIBSVM (21). The cross-validation accuracy (success rate in predicting the disease status of inputs not part of the training set) of the models generated ranged from 82.9% to 95.1% with a mean of 88.7%. The area under the curve (AUC) for the receiver operating characteristic curve was also computed. The AUC ranged from 0.89 to 0.96 with a mean of 0.93. These values for the cross-validation accuracy and the AUC suggest a good performance of the SVM models constructed.

Each model provided a weight, w_i, to the 517 SNPs ($i = 1, 2, \ldots, 517$). The absolute weight, $|w_i|$, is a measure of how important the *i*th SNP is in the model for predicting the disease status of an individual, with larger values of $|w_i|$ corresponding to more significant SNPs. We computed the mean absolute weight, $|w_i|_{avg}$, of each of the 517 SNPs as well as its standard deviation, σ_i, across the 100 models. The mean absolute weight provides a ranking of the SNPs according to the SVM method. The largest value of $|w_i|_{avg}$ across the 517 SNPs was denoted as $|w|_{max}$, and the corresponding SNP was the top-ranked SNP using the SVM method. The σ_i corresponding to this SNP was denoted as σ.

RESULTS

Genotyping findings. In the discovery cohort, an overall call rate of 95% was obtained on genome-wide genotyping using the Caucasian-based Illumina Human660W Quad BeadChip. Approximately 5% of SNPs were not called in the study cohort, and this may have been due either to assay failure or to genetic architectural differences between the European and North Indian populations. Approximately 10% of SNPs were copy number variation markers, which were removed from the analysis, leaving 559,348 analyzable SNPs.

Quality control steps. Multidimensional scaling analysis did not reveal any population stratification and showed that the North Indian population is distinct from the rest, but with an expected overlap with the Gujarati Indians in Houston, Texas HapMap population (Figure 1a). After removing copy number variation, X-chromosomal, Y-chromosomal, and mitochondrial markers and following stringent quality control, 475,771 SNPs were available for comparison in 664 RA patients and 666 controls.

To assess population substructure, multidimensional scaling analysis was again carried out on the study cohort, which demonstrated some degree of heterogeneity. After removing scattered outliers, 1 major cluster (556 patients and 590 controls) and 2 smaller subclusters (subcluster 1 [S1], 47 patients and 44 controls; subcluster 2 [S2], 61 patients and 32 controls) were observed (Figure 1b). Analysis of allele frequency across these 3 clusters indicated that <20% of the markers contributed to the substructuring in the study population (further information is available from the corresponding author). In these 2 subclusters, which had comparable numbers of patients and controls, genomic inflation was checked and no inflation within the subgroups was found ($\lambda_{S1} = 1.009$; $\lambda_{S2} = 1.01$). An association study conducted on the total sample set illustrated inflation ($\lambda = 1.25$), which increased to $\lambda = 1.31$ after removing the subclus-

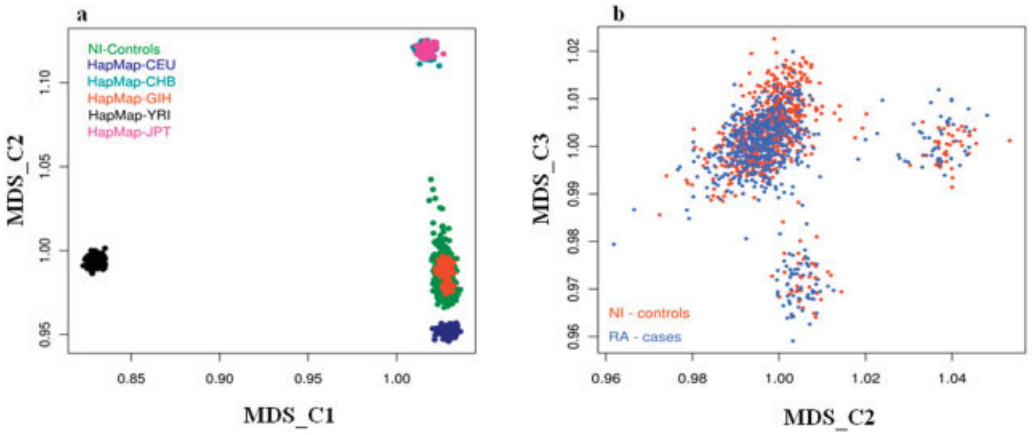

Figure 1. Multidimensional scaling (MDS) plots showing C1–C2 components for North Indian (NI) controls with HapMap populations (**a**) and C2–C3 components for North Indian rheumatoid arthritis (RA) patients and North Indian controls (**b**). CEU = Utah residents with ancestry from northern and western Europe; CHB = Han Chinese in Beijing, China; GIH = Gujarati Indians in Houston, Texas; YRI = Yoruba in Ibadan, Nigeria; JPT = Japanese in Tokyo, Japan.

ters. These small changes indicate that population substructure in this study should not have any appreciable effect on the results, and we interpret the relatively high lambda values as an indication of an elevated level of inbreeding in the Indian population at large. Therefore, these subgroups were treated separately for generating *P* values using the Cochran-Armitage trend test, and the generated *P* values were later pooled with main group *P* values using the exact-effect meta-analysis method. The overall median genotype call rate for quality-controlled SNPs was >99%.

Genome-wide association analyses. The main cluster association *P* values were adjusted for C1, C2, and C3 multidimensional scaling dimensions, as the main cluster association study showed inflation (λ = 1.25) (further information is available from the corresponding author). Subgroups as well as the main cluster were also adjusted for sex. Pooled *P* values were corrected for genomic inflation (λ_{GC} = 1.004). We plotted the observed versus the expected *P* value distribution. The Q–Q plots showed a good match between the distributions of the observed *P* values and those expected by chance ($\lambda_{GCcorrected}$ = 1.004) (Figures 2a and b). Chromosome-wide distribution of the associated SNPs is presented in Manhattan plots (Figures 3a and b). To estimate the effect of the MHC, the major

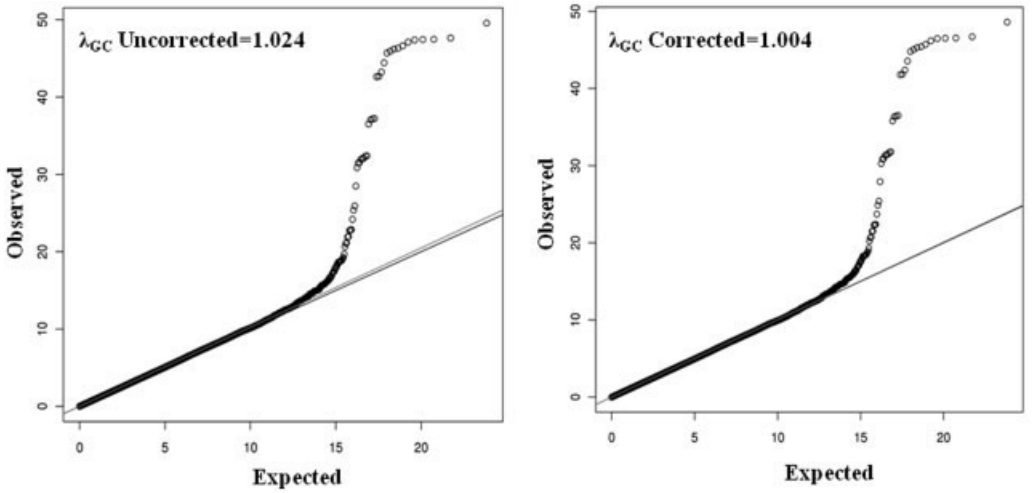

Figure 2. Q–Q plots showing the distribution of all single-nucleotide polymorphisms (~476,000) in chi-square statistical analysis and the genomic control inflation factor (λ_{GC}) in the allelic association model.

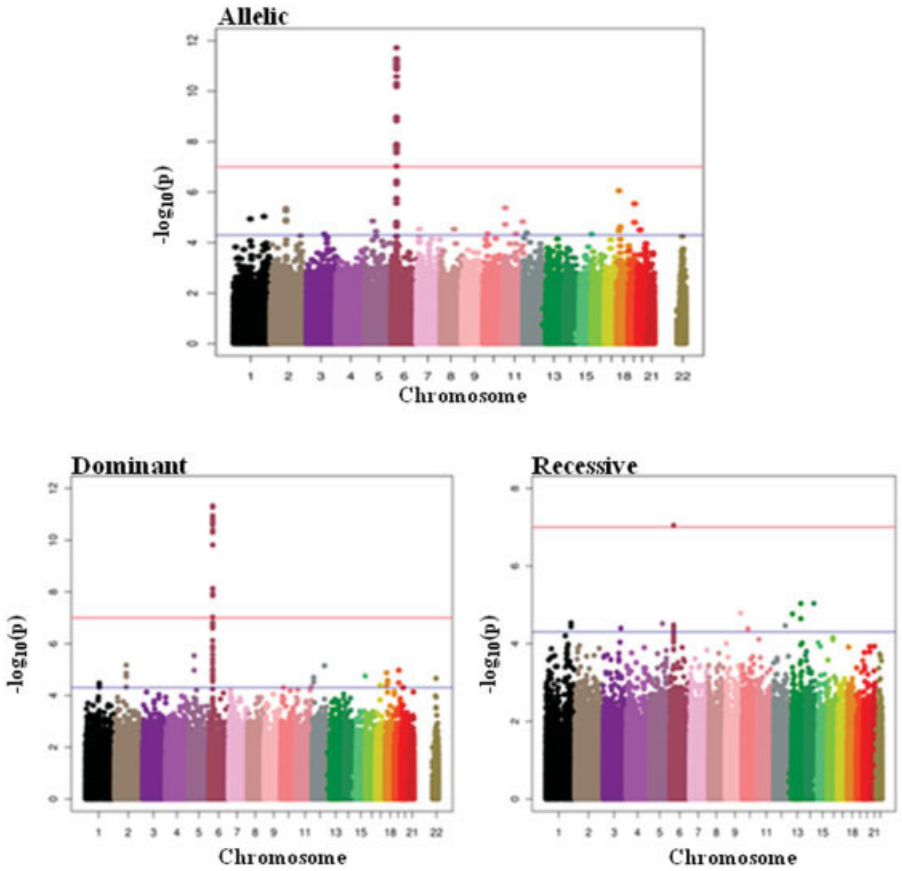

Figure 3. Manhattan plots depicting chromosomal distribution of *P* values from allelic, dominant, and recessive association models. The red line indicates the genome-wide significance cutoff. The blue line indicates the genome-wide suggestive cutoff for *P* values.

contributor to inflammatory disorders, Q–Q plots without the MHC region were plotted, which revealed that the non–HLA region genes may contribute a relatively smaller (but significant) increment to RA disease pathogenesis (further information is available from the corresponding author). To calculate the power of our GWAS stage, we used the CaTS program (http://www.sph.umich.edu/csg/abecasis/CaTS/). The GWAS stage had 80% power to detect common alleles (MAF = 0.2) that confer a genotype relative risk of 1.5 at a significance level of $P < 10^{-5}$.

A total of 27 SNPs surpassed genome-wide significance (further information is available from the corresponding author). However, all were located in the HLA region, which contributes approximately one-third of the overall genetic susceptibility to RA (22,23). Suggestive association was observed for an additional 19 genes (further information is available from the corresponding author), of which 17 were unreported (when compared to the GWAS catalog). To identify additional risk variants, association based on recessive and dominant models was carried out. Suggestive association at 17 additional novel genes was observed from these 2 models (further information is available from the corresponding author). For subsequent replication, HLA genes, LOCs (except LOC150577 and LOC730118), and a few other associated genes (due to assay limitation) were not considered.

Replication stage. In the replication stage, we genotyped 12 non-HLA markers (6 allelic top hits, 2 from dominant models and 4 from recessive models) in an independent sample set consisting of 927 RA patients and 1,148 controls. In addition, we included rs1673649 of *HLA–DQB2* as a positive control. Evidence of association ($P \leq 0.05$) was found at 4 SNPs (Table 1). To improve the power of the study, combined analysis of the GWAS and replication genotype data was carried out to test for allelic, dominant, and recessive associations. We observed suggestive evidence of association with only 1 gene, namely, ADP-ribosylation factor–like

Table 1. Replicated and novel genes/loci in the discovery and replication stages and in the combined analysis in the North Indian rheumatoid arthritis cohort, and comparison with CEU meta-analysis findings (3)*

SNP†	Chr	Chr position	Gene	Location	Alleles‡	MAF in North Indian cohort	*P* in GWAS	OR (95% CI) in GWAS	*P* in replication stage	OR (95% CI) in replication stage	*P* in combined analysis	OR (95% CI) in combined analysis	Frequency of predisposing allele in CEU meta-analysis	*P* in CEU meta-analysis	OR (95% CI) in CEU meta-analysis
rs4851269	2	100830348	LOC150577	Intron	G/A	0.46	5.52×10^{-6}	1.54 (1.27–1.86)	0.47	0.95 (0.83–1.09)	0.0007	0.83 (0.75–0.93)	0.4	6.89×10^{-5}	1.1 (1.05–1.15)
rs6542920	2	100845088	LOC150577	Intron	G/A	0.44	6.88×10^{-6}	1.53 (1.28–1.87)	0.237	1.09 (0.95–1.24)	0.0003	1.21 (1.09–1.35)	0.33	4.84×10^{-5}	1.11 (1.06–1.17)
rs1160542§	2	100832155	LOC150577	Intron	G/A	0.43	7.71×10^{-6}	0.54 (0.40–0.71)	1.00	1 (0.87–1.15)	0.0009§	0.77 (0.66–0.9)	0.46 (G)¶	1.45×10^{-6}	1.12 (1.07–1.17)
rs255758§	5	53311502	ARL15#	Intron	A/C	0.18	3.36×10^{-6}	1.92 (1.39–2.43)	0.05§	1.21 (1.00–1.47)	6.57×10^{-6}§	1.42 (1.22–1.66)	0.25	0.81	1.01 (0.93–1.1)
rs1573649**	6	32731258	HLADQB2	Coding	A/G	0.44	2.75×10^{-5}	1.46 (1.17–1.72)	0.003	1.22 (1.07–1.39)	0.000003	1.28 (1.16–1.42)	0.48 (A)¶	0.28	1.03 (0.98–1.09)
rs561041††	9	129658678	ZBTB34	Flanking 3′-UTR	A/G	0.31	9.42×10^{-6}	0.34 (0.17–0.50)	0.03††	1.42 (1.04–1.94)	0.84	0.99 (0.88–1.11)	0.28	0.90	1 (1–1)
rs4910287	11	11260163	GALNTL4	Flanking 3′-UTR	G/A	0.4	5.20×10^{-6}	0.65 (0.55–0.83)	0.40	0.94 (0.82–1.08)	0.0002§	0.75 (0.64–0.87)	0.25	0.48	0.98 (0.93–1.04)
rs1037013§	12	107219308	RIC8B	Intron	A/G	0.46	8.08×10^{-6}	1.91 (1.24–2.29)	0.42	1.05 (0.93–1.2)	0.02	1.13 (1.02–1.25)	0.43	0.84	1 (1–1)
rs7328282††	13	99554955	DOCK9	Intron	A/G	0.44	5.19×10^{-6}	2.02 (1.62–3.12)	0.03	0.86 (0.75–0.98)	0.82	1.01 (0.91–1.12)	0.4	0.05	1.05 (1–1.1)
rs2094497††	13	27910122	RASL11A	Flanking 3′-UTR	A/G	0.39	9.97×10^{-6}	2.16 (1.45–3.04)	0.58	0.96 (0.84–1.10)	0.02††	1.28 (1.04–1.56)	0.41 (A)¶	0.81	1.01 (0.93–1.1)
rs12881250††	14	95425711	LOC730118	Flanking 3′-UTR	A/C	0.29	5.15×10^{-6}	2.60 (1.65–4.09)	0.49	1.05 (0.91–1.21)	0.003	1.19 (1.06–1.33)	0.4	0.77	1.01 (0.94–1.08)
rs2002212	18	12006035	IMPA2	Intron	G/A	0.08	1.12×10^{-6}	1.96 (1.47–2.58)	0.91	0.99 (0.8–1.22)	0.001	1.30 (1.11–1.53)	0.05	0.83	1.01 (0.92–1.11)
rs9941467	19	38629027	SIPA1L3	Intron	G/A	0.1	3.59×10^{-6}	1.88 (1.38–2.46)	0.21	0.88 (0.71–1.08)	0.063	1.16 (0.99–1.36)	0.15	0.97	1 (1–1)

* SNP = single-nucleotide polymorphism; Chr = chromosome; MAF = minor allele frequency; GWAS = genome-wide association study; 3′-UTR = 3′-untranslated region.
† All unmarked rs identification numbers are from the allelic model.
‡ Second allele represents the minor allele and is the same in the discovery and replication sets in North Indians and in Utah residents with ancestry from northern and western Europe (CEU).
§ Dominant model.
¶ Allele in parentheses indicates the minor allele in CEU, which is opposite to that in North Indians.
# Association determined with 81% power.
** Used as a positive control in the replication cohort.
†† Recessive model.

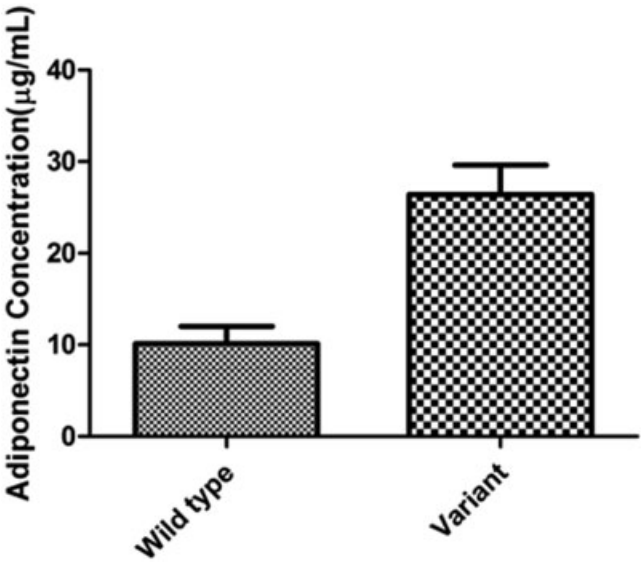

Figure 4. Adiponectin levels in individuals with wild-type (AA) and variant (CC) genotypes of rs255758 in *ARL15.* Values are the mean ± SD. A significant difference ($P < 0.0001$) in adiponectin levels was found between rheumatoid arthritis patients harboring the wild-type genotype and those harboring the variant genotype.

15 (*ARL15)*, a non–chromosome 6 gene (Table 1). To fine-map the association signals in this gene, we performed imputation-based association analysis (details are available from the corresponding author). We found suggestive association at 3 SNPs, namely, rs697109 ($P = 4.74 \times 10^{-6}$), rs697108 ($P = 4.82 \times 10^{-6}$), and rs31127 ($P = 9.28 \times 10^{-6}$). Of these, rs697109 and rs697108 statistically exceeded the effect of the landmark SNP (rs255758; $P_{\text{combined}} = 6.57 \times 10^{-6}$), the only non-HLA SNP that remained significant in both the discovery and replication stages and in the combined analysis (Table 1). However, all 4 significant *ARL15* SNPs were in strong linkage disequilibrium (LD) with each other (r^2 and $D' > 0.9$). The contribution of the risk allele (C) of rs255758 was confirmed by assessing its influence on adiponectin levels among RA patients.

Correlation of serum adiponectin levels with *ARL15* variant. In the first step toward confirming the possible role of *ARL15*, the only novel gene identified in our study, we used an enzyme-linked immunosorbent assay (KHP0041; Invitrogen) to measure the serum concentrations of adiponectin in 53 patients with RA, carrying wild-type (AA; n = 27) and variant (CC; n = 26) genotypes. Samples were prepared at the appropriate dilutions and assayed according to the manufacturer's protocol. Adiponectin levels are expressed as the mean. Statistical analysis was performed using GraphPad Prism software, and the Mann-Whitney nonparametric test was used to determine the significance of intergroup differences in serum adiponectin concentrations. The intra- and interassay coefficients of variation for adiponectin were also calculated. A significant difference ($P < 0.0001$) in adiponectin levels was found between RA patients harboring the wild-type genotype and those harboring the variant genotype (Figure 4).

Results of SVM analysis of GWAS data. The SVM method provided a ranking of the SNPs based on their mean absolute weight, $|w_i|_{\text{avg}}$. The largest value of the mean absolute weight, $|w|_{\text{max}}$, was found to be 0.572. The corresponding SNP, the top-ranked SNP according to the SVM method, was identified as rs10059065. SNPs whose mean absolute weight is close to $|w|_{\text{max}}$ may also be considered potentially significant contributors to disease susceptibility. The standard deviation, σ, is a measure of the fluctuation of the absolute weight of the top-ranking SNP across the 100 models considered. There were 114 SNPs and corresponding genes whose $|w_i|_{\text{avg}}$ was within 3σ of $|w|_{\text{max}}$ (further information is available from the corresponding author). Considering a more stringent cutoff of 1σ, we found only 6 SNPs with a $|w_i|_{\text{avg}}$ within 1σ of $|w|_{\text{max}}$. These were identified as rs10059065, rs2218970, rs17023457, rs2199998, rs1892458, and rs11605437. The corresponding genes/locus were *LOC391845*, *NRP1*, *HAO2*, *HS3ST3B1*, *LY86*, and *ETS1*, respectively. These may be considered the genes with the most significant disease association as identified by the SVM method for the present GWAS data. Most of these genes are known to be associated with other autoimmune diseases (Table 2),

Table 2. Known disease associations of top genes identified in support vector machine analysis

Gene (gene name)	Association with disease
NRP1 (neuropilin 1)	Inflammatory bowel disease, Crohn's disease, type 1 diabetes mellitus, osteonecrosis, ankylosing spondylitis
HAO2 (hydroxyacid oxidase 2)	Acquired immunodeficiency syndrome, urinary bladder neoplasms
HS3ST3B1 (heparan sulfate [glucosamine] 3-*O*-sulfotransferase 3B1)	Liver disease
LY86 (lymphocyte antigen 86)	Asthma, atherosclerosis, Crohn's disease, nasopharyngeal neoplasms, venous thromboembolism, mite-sensitive allergy
ETS1 (v-ets erythroblastosis virus E26 oncogene homolog 1 [avian])	Lupus erythematosus, systemic lupus nephritis, celiac disease, rheumatoid arthritis

and one of them, *ETS1*, was also reported to be associated with RA in 2 different studies (24,25).

We have also identified the pathways and processes in which these genes participate, as well as their known interaction partners, using NCBI resources and the databases Biogrid (http://thebiogrid.org/) (26), Reactome (http://www.reactome.org/) (27), and String (http://string-db.org/) (28) (further information is available from the corresponding author). Using the SVM method, the SNP rs31127 appears within 3σ of the highest weight SNP; this SNP corresponds to the gene *ARL15*, which, using conventional chi-square statistical analysis, we have identified as a novel gene associated with RA (further information is available from the corresponding author).

DISCUSSION

Most GWAS of RA were conducted primarily in populations with Caucasian ethnicity; a few were conducted in populations with Asian ancestry. However, it is still not known how many of the disease-associated genes/loci findings in Caucasian populations can be generalized to diverse ethnic groups. Herein we report the first GWAS of RA focused on populations of genetically distinct North Indian origin. In the discovery stage, which comprised 706 RA patients and 761 controls, 27 SNPs surpassed genome-wide significance. These were located in the HLA region (HLA–DQB1, HLA–DRA, and HLA–DQA2) (further information is available from the corresponding author), which has the most reproducible association with RA across ethnic groups (29). This unequivocally demonstrates the contribution of HLA region genes to RA susceptibility in an ethnically different North Indian population as well. Although the HLA locus has a well-established association with RA, notable differences in LD between Caucasians and North Indians (data not shown) can be exploited to identify common or population-specific causative alleles in this important locus using trans-ethnic–based fine-mapping. Since this locus is not a novel finding, we have not analyzed the association further in this study. However, replication of the association of this locus with RA susceptibility in our cohort confirms that our study was well-powered to identify common alleles of large effect.

Of the 12 non–HLA region SNPs with suggestive association that were brought forward for replication in an independent cohort, only rs255758 in *ARL15* showed stronger association in the combined analysis ($P_{\text{combined}} = 6.57 \times 10^{-6}$) (Table 1). Although this SNP did not meet the prespecified threshold for genome-wide significance, the direction of its effect was similar in both the genotyping stages and in the combined analysis.

Recently, GWAS of type 2 diabetes mellitus and coronary heart disease in subjects of European ancestry have shown pronounced association of an intronic SNP (rs4311394) in *ARL15* with circulating adiponectin levels (30). Similar findings were also reported in a meta-analysis of type 2 diabetes mellitus and metabolic traits (31). Adiponectin is one of the important adipokines that is highly expressed in the synovial fluid and synovial membrane of patients with RA (32). It affects multiple cells in the synovial tissue, such as synovial fibroblasts, chondrocytes, and osteoblasts, besides acting on inflammatory cells (33). In RA, the levels of adiponectin are raised and correlate with disease severity (34,35). Patients with a low body mass index have higher levels of adiponectin and show greater joint damage (36). This is related to increased production of proinflammatory mediators such as interleukin-6 (IL-6), matrix metalloproteinases, and prostaglandins and chemokines such as IL-8. Most of these mediators are produced by synovial fibroblasts when adiponectin binds to the adiponectin receptor on these cells and activates the NF-κB pathway (37). Given the role of adiponectin in RA, we investigated the effect of rs255758 on adiponectin levels. Our results demonstrated that the CC genotype was robustly associated with increased adiponectin levels and increased risk of RA. These observations underscore the importance of adiponectin in the pathogenesis of RA and also as a potential therapeutic target.

In addition to the conventional association analysis, we employed a machine learning approach (SVM analysis) to identify additional susceptibility SNPs/loci and their potential interacting partners. The 6 SNPs obtained after a stringent cutoff of 1σ using the SVM method corresponded to the genes/locus *LOC391845*, *NRP1*, *HAO2*, *HS3ST3B1*, *LY86*, and *ETS1*. Notably, most of these genes are also associated with inflammatory and autoimmune disorders (Table 2). For example, *NRP1* is known to be associated with IBD, Crohn's disease, and type 1 diabetes mellitus. *LY86* is reported to be associated with IBD. *ETS1* has been reported to be associated with celiac disease and was found very recently to be associated with RA in the Japanese (24) and Caucasian (25) populations. *ETS1* also appears in the list of suggestive associations in the discovery stage of the present study (further information is available from the corresponding author). *LIF*, within 2σ of $|w|_{\max}$ (further information is available from the corresponding author), has already been reported in GWAS of IBD (38,39).

This may suggest shared pathways and/or disease mechanisms between RA and other autoimmune disorders.

Within 3σ of $|w|_{max}$, we found that the top hit in our conventional association study (namely, *ARL15*) and a few genes with suggestive association (namely, *IMPA2*, *DOCK9*, *RASL11A,* and *SIPA1L3*) were also identified, using the SVM method, as potentially associated with disease (further information is available from the corresponding author). Using various publicly available databases, we have also identified the interacting partners of the top 5 genes that were determined using the SVM method (further information is available from the corresponding author). Several of these interacting partners have a known functionality relevant to disease. For example, *NRP1* interacts with *VEGFA*, which is a drug target of RA treatment. *ETS1* interacts with *JAK3* and *JAK1*, and *LY86* interacts with *CD180*, *LY96*, and *TLR4*, which play an important role in the immune system. The relevance of *NRP1*, *ETS1*, and *LY86* to disease biology is apparent (further information is available from the corresponding author). Taken together, all of these findings suggest that the SVM method is a useful tool in conjunction with GWAS.

Finally, we compared the association status of all RA index SNPs that were selected in the replication phase with European meta-analysis data (3), first, to strengthen the evidence of apparent complexity and genetic heterogeneity underlying RA pathogenesis and second, to provide a substrate for cross-ethnicity fine-mapping to characterize population-specific variation(s) relevant to pathology. Furthermore, differences in genomic architecture, effect size, and environmental factors in particular between these 2 populations may also influence detection power of the 2 scans. This has been exemplified in the present study, in which we observed that 1) 13.3% of SNPs were rare variants, unlike the case in Caucasians; 2) the strength of association of LOC150577 (rs1160542) with disease was notably different between 2 populations, even though the MAFs were comparable (North Indian MAF = 0.43; MAF of Utah residents with ancestry from northern and western Europe = 0.46) (Table 1); and 3) the association of *ARL15* (rs255758) with disease was absent in Caucasians due to differences in MAF (Table 1).

In summary, our study identified 1 novel non-HLA gene, namely, *ARL15,* in the combined analysis. This study not only advances our understanding of the genetic basis of RA but also highlights the value of performing GWAS in diverse ancestral populations. We might have missed a few novel susceptibility genes, either due to inadequate content in the currently available Caucasian LD-based SNP arrays for testing Indian populations, or due to a comparatively small sample size for detecting alleles with a minor effect.

ACKNOWLEDGMENTS

We gratefully acknowledge Ms Anjali Dabral and Ms Shalini Sharma for preparing and maintaining DNA samples. We thank Mr. Surojit Bose, LeadInvent Technologies, for unstinting computational support. We thank the Central Instrumentation facility at the University of Delhi South Campus for DNA sequencing; Sandor Proteomics for generating the discovery stage genotype data; and AceProbe Technologies for generating the replication stage genotype data.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Thelma had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. G. Juyal, A. Kumar, Jain, R. C. Juyal, Thelma.

Acquisition of data. Negi, G. Juyal, Senapati, Prasad, A. Gupta, Singh, Kashyap, A. Kumar, U. Kumar, R. Gupta, Kaur, Agrawal, Aggarwal, Jain, R. C. Juyal, Thelma.

Analysis and interpretation of data. Negi, G. Juyal, Senapati, Singh, Ott, Jain, R. C. Juyal, Thelma.

REFERENCES

1. Jarvinen P, Aho K. Twin studies in rheumatic diseases. Semin Arthritis Rheum 1994;24:19–28.
2. MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K, et al. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. Arthritis Rheum 2000;43:30–7.
3. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet 2010;42:508–14.
4. Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat Genet 2012;44:1336–40.
5. McGovern DP, Gardet A, Torkvist L, Goyette P, Essers J, Taylor KD, et al. Genome-wide association identifies multiple ulcerative colitis susceptibility loci. Nat Genet 2010;42:332–7.
6. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 2011;43:246–52.
7. McCarthy MI. Casting a wider net for diabetes susceptibility genes. Nat Genet 2008;40:1039–40.
8. Waters KM, Le Marchand L, Kolonel LN, Monroe KR, Stram DO, Henderson BE, et al. Generalizability of associations from prostate cancer genome-wide association studies in multiple populations. Cancer Epidemiol Biomarkers Prev 2009;18:1285–9.
9. Yamada H, Penney KL, Takahashi H, Katoh T, Yamano Y, Yamakado M, et al. Replication of prostate cancer risk loci in a Japanese case-control association study. J Natl Cancer Inst 2009;101:1330–6.
10. Prasad P, Kumar A, Gupta R, Juyal RC, Thelma BK. Caucasian

and Asian specific rheumatoid arthritis risk loci reveal limited replication and apparent allelic heterogeneity in North Indians. PLoS One 2012;7:e31584.
11. Freudenberg J, Lee HS, Han BG, Shin HD, Kang YM, Sung YK, et al. Genome-wide association study of rheumatoid arthritis in Koreans: population-specific loci as well as overlap with European susceptibility loci. Arthritis Rheum 2011;63:884–93.
12. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
13. Yu K, Wang Z, Li Q, Wacholder S, Hunter DJ, Hoover RN, et al. Population substructure and control selection in genome-wide association studies. PLoS One 2008;3:e2551.
14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007; 81:559–75.
15. Boser BE, Guyon IM, Vapnik VN. Training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory; 1992 July 27-29; Pittsburgh, Pennsylvania. New York: ACM Press; 1992. p. 144–52.
16. Cortes C, Vapnik V. Support vector networks. Machine Learn 1995;20:273–97.
17. Waddell M, Page D, Zhan F, Barlogie B, Shaughnessy J Jr. Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma. BIOKDD'05: Proceedings of the 5th International Workshop on Bioinformatics; August 21-4, 2005; Chicago, Illinois. New York: ACM Press; 2005. p. 21–8.
18. Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, et al. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet 2009;5:e1000678.
19. Roshan U, Chikkagoudar S, Wei Z, Wang K, Hakonarson H. Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. Nucleic Acids Res 2011;39:e62.
20. Mittag F, Buchel F, Saad M, Jahn A, Schulte C, Bochdanovits Z, et al, for the International Parkinson's Disease Genomics Consortium (IPDGC). Use of support vector machines for disease risk prediction in genome-wide association studies: concerns and opportunities. Hum Mutat 2012;33:1708–18.
21. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2011;2:1–27.
22. Deighton CM, Walker DJ, Griffiths ID, Roberts DF. The contribution of HLA to rheumatoid arthritis. Clin Genet 1989;36: 178–82.
23. Rigby AS, Silman AJ, Voelm L, Gregory JC, Ollier WE, Khan MA, et al. Investigating the HLA component in rheumatoid arthritis: an additive (dominant) mode of inheritance is rejected, a recessive mode is preferred. Genet Epidemiol 1991;8:153–75.
24. Okada Y, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, et al. Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. Nat Genet 2012;44:511–6.
25. Chatzikyriakidou A, Voulgari PV, Georgiou I, Drosos AA. Altered sequence of the ETS1 transcription factor may predispose to rheumatoid arthritis susceptibility. Scand J Rheumatol 2013;42: 11–4.
26. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, et al. The BioGRID interaction database: 2011 update. Nucleic Acids Res 2011;39:D698–704.
27. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, et al. Reactome knowledgebase of biological pathways and processes. Nucleic Acids Res 2008;37:D619–22.
28. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: functional interaction networks of proteins globally integrated and scored. Nucleic Acids Res 2011;39:D561–8.
29. Fernando MM, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, et al. Defining the role of the MHC in autoimmunity: a review and pooled analysis. PLoS Genet 2008;4:e1000024.
30. Richards JB, Waterworth D, O'Rahilly S, Hivert MF, Loos RJ, Perry JR, et al, and GIANT Consortium. A genome-wide association study reveals variants in ARL15 that influence adiponectin levels. PLoS Genet 2009;5:e1000768.
31. Dastani Z, Hivert MF, Timpson N, Perry JR, Yuan X, Scott RA, et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. PLoS Genet 2012;8:e1002607.
32. Tan W, Wang F, Zhang M, Guo D, Zhang Q, He S. High adiponectin and adiponectin receptor 1 expression in synovial fluids and synovial tissues of patients with rheumatoid arthritis. Semin Arthritis Rheum 2009;38:420–7.
33. Frommer KW, Zimmermann B, Meier FM, Schroder D, Heil M, Schaffler A, et al. Adiponectin-mediated changes in effector cells involved in the pathophysiology of rheumatoid arthritis. Arthritis Rheum 2010;62:2886–99.
34. Klein-Wieringa IR, van der Linden MP, Knevel R, Kwekkeboom JC, van Beelen E, Huizinga TW, et al. Baseline serum adipokine levels predict radiographic progression in early rheumatoid arthritis. Arthritis Rheum 2011;63:2567–74.
35. Giles JT, van der Heijde DM, Bathon JM. Association of circulating adiponectin levels with progression of radiographic joint destruction in rheumatoid arthritis. Ann Rheum Dis 2011;70: 1562–8.
36. Baker JF, George M, Baker DG, Toedter G, Von Feldt JM, Leonard MB. Associations between body mass, radiographic joint damage, adipokines and risk factors for bone loss in rheumatoid arthritis. Rheumatology (Oxford) 2011;50:2100–7.
37. Gomez R, Conde J, Scotece M, Gomez-Reino JJ, Lago F, Gualillo O. What's new in our understanding of the role of adipokines in rheumatic diseases? Nat Rev Rheumatol 2011;7:528–36.
38. Franke A, Hampe J, Rosenstiel P, Becker C, Wagner F, Hasler R, et al. Systematic association mapping identifies NELL1 as a novel IBD disease gene. PLoS ONE 2007;8:e691.
39. Imielinski M, Baldassano RN, Griffiths A, Russell RK, Annese V, Dubinsky M, et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat Genet 2009;4: 1335–40.