Pathway analysis of genome-wide association studies on rheumatoid arthritis

G.G. Song[1], S.-C. Bae[2], Y.H. Lee[1]

[1]Division of Rheumatology, Department of Internal Medicine, Korea University College of Medicine; [2]Hospital for Rheumatic Diseases, Hanyang University Medical Centre, Seoul, Korea.

Abstract

Objectives

The aims of this study were to identify candidate single nucleotide polymorphisms (SNPs) and candidate mechanisms of RA and generate hypotheses for SNP → gene→ pathways.

Methods

We used a meta-analysis dataset of rheumatoid arthritis (RA) genome-wide association studies (GWAS) which included 2,554,714 SNPs in 5,539 RA cases and 20,169 controls of European descent. ICSNPathway (Identify candidate Causal SNPs and Pathways) analysis was applied to the meta-analysis results of the RA GWAS dataset.

Results

ICSNPathway analysis identified 49 candidate SNPs included in 37 candidate pathways. The top 5 candidate causal SNPs, rs1063478 (p=5.40E-09), rs 375256 (p=3.44E-09), rs365066 (p=3.60E-30), rs2581 (p=2.7E-25), and rs1059510 (p=2.52E-06) were all at human leukocyte antigen (HLA) loci. These candidate SNPs and pathways provided 22 hypothetical biological mechanisms. The most strongly associated pathway concerned HLA: rs1063478 alters the role of HLA-DMA in the context of the pathway of antigen processing and presentation of peptide antigen. ICSNPathway analysis identified two candidate non-HLA SNPs included in ten candidate pathways, which provided two hypothetical biological mechanisms. First, rs2476601 alters the role of protein tyrosine phosphatase non-receptor 22 (PTPN22) in the context of immune response-activation cell surface receptor signalling pathway, and, rs2230926 alters the role of tumour necrosis factor-alpha-induced protein 3 (TNFAIP3) in the context of the CD40L signalling pathway.

Conclusion

The application of ICSNPathway analysis to the meta-analysis results of RA GWAS datasets indicated candidate SNPs and pathways involving HLA-DMA, PTPN22, and TNAIP3 associated with RA susceptibility.

Key words

rheumatoid arthritis, GWAS, meta-analysis, pathway-based analysis

Gwan Gyu Song, MD, PhD
Sang-Cheol Bae, MD, PhD
Young Ho Lee, MD, PhD

Please address correspondence and reprint requests to:
Young Ho Lee, MD, PhD,
Division of Rheumatology,
Department of Internal Medicine,
Korea University Anam Hospital,
Korea University College of Medicine,
126-1 5 ga, Anam-dong,
Seongbuk-gu,
Seoul 136-705, Korea.
E-mail: lyhcgh@korea.ac.kr

Received on November 20, 2012; accepted in revised form on January 29, 2013.

Funding: This study was supported by a grant from the Korea Healthcare Technology R&D Project, Ministry of Health and Welfare, Republic of Korea (A102065).

Competing interests: none declared.

Introduction

Rheumatoid arthritis (RA) is a chronic inflammatory disease of predominantly synovial joints that affects up to 1% of the world's population (1, 2). Although the aetiology of RA has not been determined, a genetic component of susceptibility to RA has been established by twin and family studies, which estimated that the heritability of RA liability may be as high as 60% (3). Human leukocyte antigen (HLA) class II molecules are the most powerful genetic factors of RA identified to date, but family studies suggest that this association accounts for only one-third of genetic susceptibility, and that non-HLA genes are also involved. Genome-wide association studies (GWAS) offer a powerful means of searching for genes that confer susceptibility to complex diseases (4). As a result, the number of GWAS being reported is growing rapidly and this has led to the discovery and replication of new disease genes (5). However, although large-scale GWAS have been carried out on complex diseases, including RA, much of the genetic component of variation in RA remains unexplained.

The increased availability of GWAS datasets provides powerful research opportunities. Although RA GWAS data have shown that the HLA region on chromosome 6p plays a key role in RA susceptibility, other genes also appear to account for the genetic contribution to RA susceptibility (6). It appears that individual genes and genetic variants make small risk contributions by interacting with each other to cause RA. However, genetic signals have been examined at the single marker level in RA GWAS studies, and the biological mechanisms identified are controversial. One of the key challenges of GWAS data interpretation is to identify causative SNPs and to provide evidence and hypothetical mechanisms responsible for observed traits (7). Thus, we considered that the using of new methods to study existing GWAS data might provide additional biological insights and highlight new candidate gene. ICSNPathway (Identify candidate Causal SNPs and Pathways) analysis was developed to identify candidate causal SNPs and their corresponding candidate causal pathways from GWAS data by integrating linkage disequilibrium (LD) analysis, functional SNP annotation, and pathway-based analysis (PBA) (8). It combines the analysis of candidate SNPs and PBA to generate hypothesis of SNP → gene → pathways which indicates that the candidate SNP alters the role of its corresponding gene/protein in the context of the pathway associated with traits (8).

Pathway analysis using meta-analysis dataset may increase more statistical power than analysis using individual data, because meta-analysis of GWAS datasets increases power to detect association signals by increasing sample size and by allowing the examination of more variants than individual datasets (9). Thus, the objective of this study was to identify candidate SNPs and candidate mechanisms in RA, and to generate SNP to gene to pathway hypotheses associated with RA, by applying ICSNPathway analysis to the meta-analysis data of RA GWAS datasets.

Materials and methods

Study populations

The meta-analysis results of RA GWAS datasets were used as the data was publicly available at http://www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_*et al*._2010NG/, and originated from the study conducted by Stahl *et al*. (6), which included 2,554,714 SNPs in 5,539 autoantibody-positive RA cases and 20,169 controls of European descent from 6 RA GWAS datasets. Details of the RA GWAS meta-analysis performed are included in the Stahl *et al*. paper in Supplementary Fig. 1.

Identification of candidate causal SNPs and pathways

ICSNPathway analysis was applied to the RA GWAS meta-analysis results. ICSNPathway analysis was conducted in two stages (8). The first involved the pre-selection of candidate SNPs by LD analysis and functional SNP annotation based on the most significant SNPs in the GWAS. The second stage involved the annotation of biological mechanisms to the pre-selected candidate causal SNPs by using the PBA

algorithm, a process referred to as improved-gene set enrichment analysis (*i*-GSEA). A full list of GWAS SNP *p*-values was inputted for ICSNPathway analysis. There are two key concepts and one key algorithm applied in ICSNPathway. One concept is in LD analysis, which searches the SNPs in LD with the most significant SNPs of GWAS to ensure to capture more possible candidate causal SNPs based on the extended data set which includes HapMap data. All SNPs in HapMap were included in the first stage. The other concept is functional SNPs. ICSNPathway pre-selects candidate causal SNPs based on functional SNPs, which are important for understanding the underlying genetics of human health. Functional SNPs are defined as SNPs that may alter protein, gene expression or the role of protein in context of pathway. The functional SNPs include deleterious and non-deleterious non-synonymous SNPs, SNPs leading in gain or loss of stop codon, SNPs resulting in frame shift, SNPs in essential splice site and SNPs in regulatory region.

The ICSNPathway server implements a PBA algorithm, as named improved-gene set enrichment analysis (*i*-GSEA), on the full list of GWAS SNP *p*-values to detect pathways associated with traits. Briefly, 1) each SNP is mapped to its nearest gene according to the SNP and gene localisation in Ensembl 61 database (http://www.wnsembl.org/biomart/martview), and the maximum t = -log (*p*-value) of the SNPs mapped to a gene is assigned to represent the gene. Then all the genes are raked by decreasing their representation value t (2). For each pathway S, ES (enrichment score, *i.e.* a Kolmogor-Smirnov like running-sum statistics with weight [a]) which measures the tendency that genes of a pathway are located at the top of the ranked gene list, is calculated (3). ES is converted to SPES (significant proportion based enrichment score) by multiplying it to m_1/m_2, where m_1 is the proportion of significant genes (defined as genes mapped with at least one of the top 5% most significant SNPs of all SNPs in GWAS) for pathways S, and m_2 is the proportion of significant genes for all the genes in the GWAS (4). SNP label permutation and normalisation are employed to generate the distribution of SPES and to correct gene variation (the bias due to different genes with different number of mapped SNPs) (5). Based on all the distribution of SPESs generated by permutation, nominal *p*-value is calculated and false discovery rate (FDR) is computed for multiple testing correction (10). By "the most significant SNPs", it is meant that SNPs with *p*-value below certain threshold. We can specify the *p*-value threshold to extract the most significant SNPs from the GWAS SNP *p*-values. ICSNPathway drew significant pathways from the original GWAS when we chose *p*-value ($<10^{-5}$) threshold. Thus, we used *p*-value ($<10^{-5}$) as *p*-value threshold.

Two parameters were set for this analysis. The first was 'within gene,' meaning that only *p*-values of SNPs located within genes were utilised in the PBA algorithm, and the second was a false discovery rate cutoff (0.05) for multiple testing correction. There were no specific criteria to select gene numbers. We used the cut-off of minimum 5 and maximum 100 to avoid the overly narrow or overly broad functional categories. To avoid stochastic bias or the inclusions of general biological process, we discarded pathways that contained <5 or >100 genes. Although several options are available for the annotation of pathways, we chose the Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/pathway.html) (11), BioCarta (http://www.biocarta.com/genes/index.asp) [12], Gene ontology (GO) (http://www.geneontolog.org) (level 4 GO terms of biological process domain and molecular function domain) (12), and MSigDB (http://www.broadinstitute.org/gsea/msigdb/index.jsp) (curated GO terms of biological process domain and molecular function domain) to ensure comprehensive coverage of pathways and to obtain high-quality information for well-defined pathways.

The HLA region encodes proteins of classical HLA class I and II genes in major histocompatibility complex (HLA) and is essential for immune recognition. This region is highly polymorphic and its LD extends across HLA and non-HLA genes in the HLA, and thus, this region could influence pathway analysis. Several non-HLA loci located in the HLA region had variants reported as potential causal variants. And it remains unclear how many independent effects actually reside in the HLA region in RA. The HLA region is a large area of strong LD. When HLA-associated autoimmune diseases like RA are studied, it is important to adjust for influences from the HLA region, given its LD patterns. Therefore, the two analyses were performed, with and without the HLA region. We defined the HLA region in this study as the region on chromosome 6, from base pair 20,000,000 to base pair 40,000,000.

When a candidate SNP was not present on a particular genotyping array, proxy SNPs in LD with that candidate SNP were identified based on observed LD patterns in HapMap. Thus, SNAP, a tool used for the identification and annotation of proxy SNPs using HapMap, was used (13).

Results

Candidate SNPs and pathways resulting from the meta-analysis of RA GWAS

ICSN Pathway analysis identified 49 candidate SNPs included in 37 candidate pathways by utilising the 2.554, 714 GWAS SNP *p*-values as input and the most significant SNPs ($p<1 \times 10^{-5}$), (Tables I, II and Fig. 1). The top 5 candidate causal SNPs were rs1063478 (*p*=5.40E-09), rs375256 (*p*=3.44E-09), rs365066 (*p*=3.60E-30), rs2581 (*p*=2.7E-25), and rs1059510 (*p*=2.52E-06), which are all on HLA loci. All of the top 5 candidate SNPs, except rs1059510, were not in LD with any SNP and had genome-wide significance. SNP rs1059510, which was not represented in the original GWAS meta-analysis, is in LD with rs2252745 (r^2=0.92), which did not reach genome-wide significance in the original GWAS meta-analysis (*p*=2.52E-06). Biological mechanisms represent that the candidate SNP may alter the role of its corresponding gene/protein in the context of the pathway(s) associated with traits. These 49 candidate SNPs included in 37 candidate pathways indicated 22 hypothetical

Table I. Candidate SNPs of RA identified in the pathway analysis.

Candidate causal SNP	Functional class	Gene	Candidate causal pathway*	-log$_{10}$(p)†	In LD with	r^2	D'	-log$_{10}$(p)‡
rs1063478	non-synonymous coding	HLA-DMA	1 2 10 11 15 22 23 24 33	8.268	rs1063478	–	–	8.268
rs375256	non-synonymous coding	HLA-DOA	1 2 10 11 23	8.463	rs375256	–	–	8.463
rs365066	non-synonymous coding	HLA-DOA	1 2 10 11 23	29.444	rs365066	–	–	29.444
rs2581	regulatory region	HLA-DOA	1 2 10 11 23	24.559	rs2581	–	–	24.559
rs1059510	non-synonymous coding (deleterious)	HLA-E	3	-	rs2252745	0.92	1.0	5.599
rs2735059	non-synonymous coding	HLA-F	3	6.682	rs2735059	–	–	6.682
rs2072895	non-synonymous coding&splice site	HLA-F	3	6.676	rs2072895	–	–	6.676
rs915669	regulatory region	HLA-G	3	5.209	rs915669	–	–	5.209
rs915668	regulatory region	HLA-G	3	10.606	rs915668	–	–	10.606
rs1063320	regulatory region	HLA-G	3	10.740	rs1063320	–	–	10.740
rs3763366	regulatory region	TAP2	3 9 27	23.851	rs3763366	–	–	23.851
rs4148869	regulatory region	TAP2	3 9 27	23.282	rs4148869	–	–	23.282
rs4148876	non-synonymous coding	TAP2	3 9 27	17.646	rs4148876	–	–	17.646
rs241448	stop_lost	TAP2	3 9 27	6.772	rs241448	–	–	6.772
rs241447	non-synonymous coding	TAP2	3 9 27	6.857	rs241447	–	–	6.857
rs16870908	non-synonymous coding (deleterious)	TAP2	3 9 27	37.236	rs16870908	–	–	37.236
rs2071888	non-synonymous coding (deleterious)	TAPBP	3 9	10.117	rs2071888	–	–	10.117
rs1041981	non-synonymous coding	LTA	4 17	3.867	rs2071592	0.917	0.957	7.156
rs2229699	non-synonymous coding	LTB	4 5 19	-	rs12215563	1.0	1.0	6.472
rs4645843	non-synonymous coding (deleterious)	TNF	6 8 12 17 19 21 28 32 35	-	rs6903496	1.0	1.0	6.201
rs6472	non-synonymous coding	CYP21A2	7	18.419	rs6472	–	–	18.419
rs6474	frameshift coding	CYP21A2	7	108.359	rs6474	–	–	108.359
rs7887	non-synonymous coding	EHMT2	13	67.967	rs7887	–	–	67.967
rs2157678	regulatory region	TRIM15	14	1.788	rs1015466	0.837	1.0	5.921
rs929156	non-synonymous coding	TRIM15	14	27.979	rs929156	–	–	27.979
rs8192583	non-synonymous coding	NOTCH4	14 29	13.318	rs8192583	–	–	13.318
rs8192585	non-synonymous coding	NOTCH4	14 29	33.403	rs8192585	–	–	33.403
rs8192579	non-synonymous coding (deleterious)	NOTCH4	14 29	13.350	rs8192579	–	–	13.350
rs8192591	non-synonymous coding	NOTCH4	14 29	10.395	rs8192591	–	–	10.395
rs397081	regulatory region	NOTCH4	14 29	19.900	rs397081	–	–	19.900
rs422951	non-synonymous coding (deleterious)	NOTCH4	14 29	12.648	rs422951	–	–	12.648
rs915894	stop gained	NOTCH4	14 29	7.452	rs915894	–	–	7.452
rs2476601	non-synonymous coding (deleterious)	PTPN22	16 20 23	70.206	rs2476601	–	–	70.206
rs805299	regulatory region	APOM	18	56.750	rs805299	–	–	56.750
rs9332739	non-synonymous coding	C2	25 26 30	13.510	rs9332739	–	–	13.510
rs4151667	non-synonymous coding (deleterious)	CFB	25	14.203	rs4151667	–	–	14.203
rs1057373	regulatory region	TAP1	27	5.517	rs1057373	–	–	5.517
rs2071543	non-synonymous coding (deleterious)	PSMB8	31	27.595	rs2071543	–	–	27.595
rs241419	non-synonymous coding (deleterious)	PSMB9	31	15.297	rs241419	–	–	15.297
rs28399977	non-synonymous coding (deleterious)	MSH5	34	-	rs9267536	1.0	1.0	5.876
rs707938	non-synonymous coding	MSH5	34	5.129	rs707938	–	–	5.129
rs2607015	non-synonymous coding	VARS	36	49.924	rs2607015	–	–	49.924
rs437179	non-synonymous coding	SKIV2L	37	20.975	rs437179	–	–	20.975
rs3911893	non-synonymous coding	SKIV2L	37	7.740	rs3911893	–	–	7.740
rs449643	non-synonymous coding (deleterious)	SKIV2L	37	17.614	rs449643	–	–	17.614
rs438999	non-synonymous coding	SKIV2L	37	38.775	rs438999	–	–	38.775
rs106287	non-synonymous coding (deleterious)	SKIV2L	37	26.281	rs106287	–	–	26.281
rs2071596	non-synonymous coding	BAT1	37	14.271	rs2071596	–	–	14.271
rs2523512	non-synonymous coding	BAT1	37	8.697	rs2523512	–	–	8.697

SNP: single nucleotide polymorphism; RA: rheumatoid arthritis; LD: linkage disequilibrium.

*Numbers indicate the indexes of pathways, which are ranked by statistical significance (false discovery rate).

†-log$_{10}$(p) for candidate causal SNP in original genome-wide association studies (GWAS). '–' denotes that this SNP was not represented in the original GWAS meta-analysis.

‡-log$_{10}$(p) for the SNP (which the candidate causal SNP is in LD with) in the original GWAS meta-analysis.

ICSNPathway analysis identified 49 candidate SNPs included in 37 candidate pathways. The top 5 candidate causal SNPs were rs1063478 (p=5.40E-09), rs 375256 (p=3.44E-09), rs365066 (p=3.60E-30), rs 2581 (p=2.7E-25), and rs1059510 (p=2.52E-06), which are all on HLA loci.

Table II. Candidate pathways of RA identified in the pathway analysis.

Index	Candidate causal pathway	Description	Nominal *p*-value	FDR
1	Antigen processing and presentation of exogenous peptide antigen	GO:0002478	<0.001	<0.001
2	Antigen processing and presentation of peptide antigen via HLA class II	GO:0002495	<0.001	<0.001
3	Antigen processing and presentation of peptide antigen via HLA class I	GO:0002474	<0.001	<0.001
4	Lymph node development	GO:0048535	<0.001	<0.001
5	Interleukin-12 biosynthetic process	GO:0042090	<0.001	<0.001
6	Positive regulation of translational initiation	GO:0045948	<0.001	0.001
7	C21-Steroid hormone biosynthetic process	GO:0006700	<0.001	0.002
8	mspPathway	Msp/Ron receptor signalling pathway	0.003	0.002
9	HLA protein binding	GO:0042287	0.001	0.004
10	Regulation of lymphocyte of differentiation	GO:0045619	<0.001	0.005
11	Regulation of lymphocyte activation	GO:0051249	<0.001	0.010
12	Positive regulation of cellular biosynthetic process	GO:0031328	<0.001	0.010
13	Spermatid differentation	GO:0048515	<0.001	0.011
14	Cell_Fate_commitment	GO:0045165	<0.001	0.012
15	Positive regulation of lymphocyte differentiation	GO:0045621	0.001	0.015
16	Immune response-activating cell surface receptor signalling pathway	GO:0002429	<0.001	0.015
17	cytokinePathway	Cytokine network cytokine network	0.009	0.016
18	Lipid transporter activity	GO:0005319	<0.001	0.021
19	Positive regulation of cytokine biosynthetic process	GO:0042108	0.001	0.027
20	Immune response-regulating signalling pathway	GO:0002764	0.002	0.028
21	il10Pathway	IL-10 anti-inflammatory signalling pathway	0.005	0.032
22	Lymphocyte mediated immunity	GO:0002449	<0.001	0.032
23	Lymphocyte differentiation	GO:0030098	<0.001	0.032
24	Adaptive immune response	GO:0002460	0.001	0.032
25	Complement activation, alternative pathway	GO:0006957	0.004	0.034
26	lectinPathway	Lectin induced complement pathway	0.003	0.035
27	hsa05340	Primary immunodeficiency	0.003	0.035
28	hsp27Pathway	Stress induction of HSP Regulation	0.009	0.036
29	hsa04330	Notch signalling pathway	<0.001	0.038
30	Adaptive immune response	GO:0002250	0.005	0.039
31	Negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	GO:0051436	0.007	0.039
32	Positive regulation of translation	GO:0045727	0.016	0.040
33	Positive regulation of lymphocyte activation	GO:0051251	0.001	0.042
34	Meiotic cell cycle	GO:0051321	0.002	0.043
35	Negative regulation of cytokine secretion	GO:0050710	0.027	0.044
36	Ligase activity forming carbon oxygen bonds	GO:0016875	0.008	0.044
37	RNA helicase activity	GO:0003724	0.017	0.049

RA: rheumatoid arthritis; FDR: false discovery rate.
ICSNPathway analysis identified 49 candidate SNPs included in 37 candidate pathways. The pathway analysis does not replace the GWAS meta-analysis results, but plays a complementary part in identifying novel candidate genes or sets of genes. The results for pathway association approaches may lead to the formulation of new hypotheses for additional validations.

biological mechanisms. For example, rs1063478 alters the role of HLA-DMA in the context of the pathways of antigen processing and presentation of exogenous peptide antigen, antigen processing and presentation of peptide antigen via HLA class II, regulation of lymphocyte differentiation, regulation of lymphocyte activation, positive regulation of lymphocyte differentiation, lymphocyte mediated immunity, lymphocyte differentiation, adaptive immune response, and positive regulation of lymphocyte activation. Rs375256, rs365066, rs2581→HLA-DOA→antigen processing and presentation of exogenous peptide antigen, antigen processing and presentation of peptide antigen via HLA class II, regulation of lymphocyte differentiation, regulation of lymphocyte activation, and lymphocyte differentiation. Rs1059510→HLA-E→antigen processing and presentation of peptide antigen via HLA class I (Tables I, II).

Candidate SNPs and pathways after excluding HLA region

We additionally performed the pathway analysis without HLA region to adjust for influences from the HLA region, given its LD patterns. ICSNPathway analysis identified two candidate non-HLA SNPs included in ten candidate pathways (Tables III, IV and Fig. 1). SNP rs2476601, which had genome-wide significance in the original GWAS meta-analysis, is not in LD with any

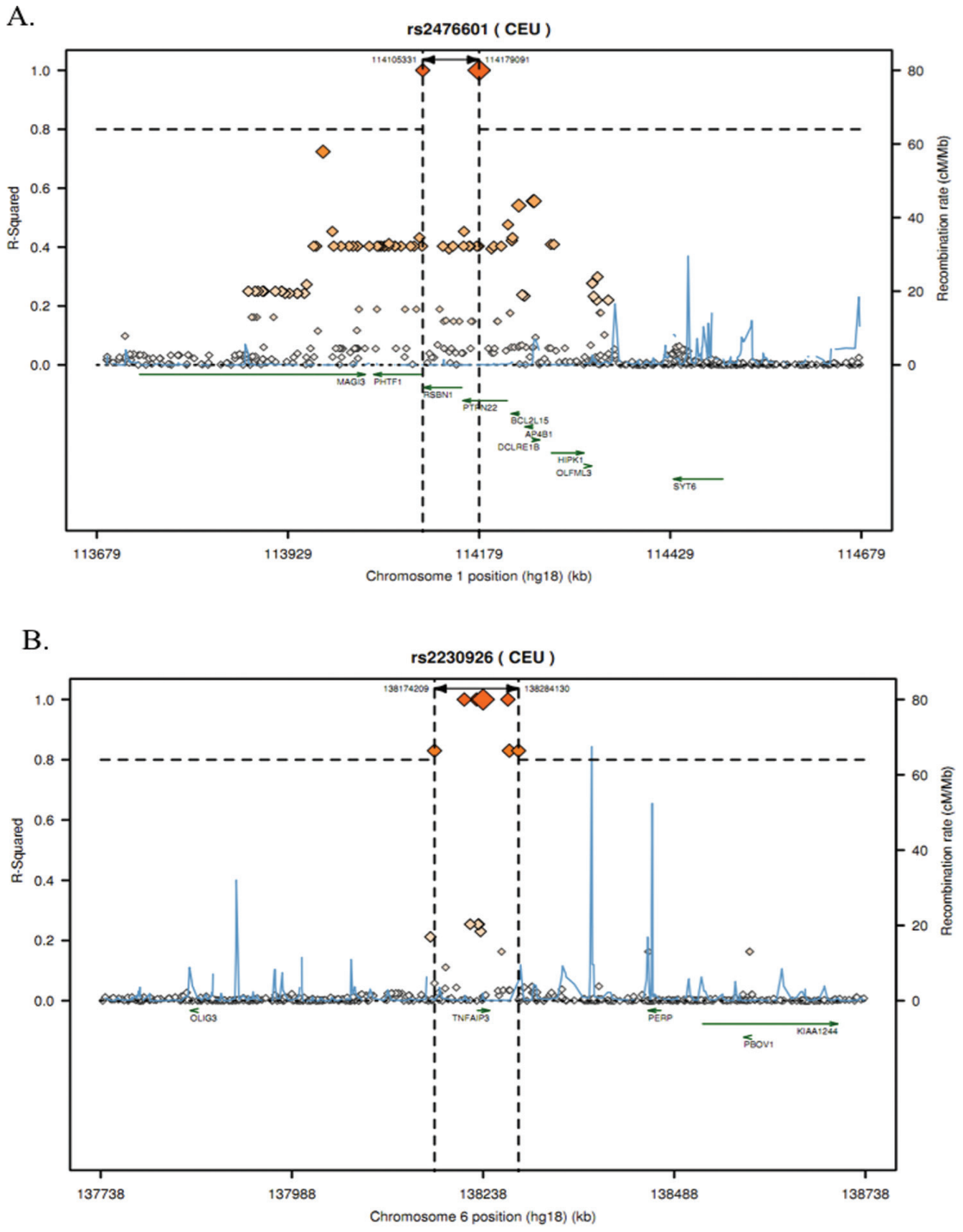

Fig. 1. Regional LD plots of the rs2476601 (PTPN22) (**A**) and rs2230926 (TNFAIP3) (**B**) SNPs. SNPs are plotted along with their proxies (based on HAPMAP CEU) as a function of genomic location, annotated by the recombination rate across the locus (light-blue line). On the y-axis, pairwise r^2 values are given for each proxy SNP using colour codes.

Table III. Candidate SNPs of RA after excluding HLA region.

Candidate causal SNP	Functional class	Gene	Candidate causal pathway*	-$\log_{10}$(p) †	In LD with	r^2	D'	-$\log_{10}$(*p*-value)‡
rs2476601	non-synonymous coding (deleterious)	PTPN22	1 2 3 5 7 8 10	70.206	rs2476601	–	–	70.206
rs2230926	non-synonymous coding	TNFAIP3	4 6 9	–	rs9494895	1.0	1.0	6.900

SNP: single nucleotide polymorphism; HLA: human leukocyte antigen; RA: rheumatoid arthritis; LD: linkage disequilibrium.*Numbers indicate the indexes of pathways, which are ranked by statistical significance (false discovery rate). †-$\log_{10}$(p) for candidate causal SNP in original genome-wide association studies (GWAS). '–' denotes that this SNP was not represented in the original GWAS meta-analysis. ‡-$\log_{10}$(p) for the SNP (which the candidate causal SNP is in LD with) in the original GWAS meta-analysis.ICSNPathway analysis identified two candidate SNPs included in ten candidate pathways after excluding HLA region.

Table IV. Candidate pathways of RA after excluding HLA region.

Index	Candidate causal pathway	Description	Nominal *p*-value	FDR
1	Immune response-activating cell surface receptor signalling pathway	GO:0002429	<0.001	<0.001
2	Immune response-regulating signalling pathway	GO:0002764	<0.001	<0.001
3	Lymphocyte differentiation	GO:0030098	<0.001	<0.001
4	Regulation of I kappaB kinase NF kappaB cascade	GO:0043122	<0.001	<0.001
5	Protein tyrosine phosphatase activity	GO:0004725	<0.001	<0.001
6	I kappaB kinase NF kappaB cascade	GO:0007249	0.001	<0.001
7	Protein amino acid dephosphorylation	GO:0006470	0.006	0.011
8	Dephosphorylation	GO:0016311	0.008	0.018
9	cd40Pathway	CD40L signalling pathway	0.013	
	Genes and SNPs in the pathway	CD40L signalling pathway		0.021
10	Phosphoprotein phosphatase activity	GO:0004721	0.015	0.022

HLA: human leukocyte antigen; RA: rheumatoid arthritis; FDR: false discovery rate.
ICSNPathway analysis identified two candidate SNPs included in ten candidate pathways after excluding HLA region. The pathway analysis does not replace the GWAS meta-analysis results, but plays a complementary part in identifying novel candidate genes or sets of genes. The results for pathway association approaches may lead to the formulation of new hypotheses for additional validations.

SNP (*p*=6.22E-71) (Fig. 1). Rs2230926, which was not represented in the original GWAS meta-analysis, is in LD with rs9494895 (r^2=1.0), which did not reach genome-wide significance in the original GWAS (*p*=1.26E-07) (Fig. 1). Proxy SNPs in LD with that candidate SNP were identified based on observed LD patterns in HapMap. SNPs are plotted along with their proxies (based on HAPMAP CEU) as a function of genomic location, annotated by the recombination rate across the locus (light-blue line) in Figure 1.

These two candidate non-HLA SNPs identified in ten candidate pathways provided two hypothetical biological mechanisms. First, rs2476601 alters the role of protein tyrosine phosphatase non-receptor 22 (PTPN22) in the context of the pathways of immune response-activation cell surface receptor signalling pathway, immune response-regulating signalling pathway, lymphocyte differentiation, protein tyrosine phosphatase activity, protein amino acid dephosphorylation, dephosphorylation, phosphoprotein phosphatase activity. Second, rs2230926 alters the role of tumour necrosis factor, alpha-induced protein 3 (TNFAIP3) in the context of the pathways of regulation of I_kappaB kinase NF kappaB cascase, I kappaB kinase NF kaappaB cascase, and the CD40L signalling pathway (Tables III and IV). Other causal variants of the TNFAIP3 have been described that are not within the coding region. One option used for this analysis was 'within gene', meaning that *p*-values of SNPs located within genes were utilised in the PBA algorithm. The PTPN22 result identified in this pathway analysis is not novel and has been known since before the GWAS era began.

Discussion

GWAS have been used successfully to identify novel common genetic variants that contribute to susceptibility to complex diseases (4). However, individual GWAS are limited in terms of the identification of new loci, because a limited set of variants are genotyped, and because the reported variant is unlikely to be the causal variant, rather it is more likely to be in LD with relevant variants. Reported loci are those that reach a certain stringent statistical "genome-wide" significance criterion, whereas hundreds of thousands of other genotyped markers have received little attention. However, multiple related genes in the same pathway may work together to confer disease susceptibility, and some of these genes may not reach genome-wide significance in any single GWAS. Thus, combined analysis of GWAS, extremely large GWAS, or PBA is required to identify new loci that leading to susceptibility to complex diseases (9, 14). Furthermore, combining results from multiple GWAS datasets may strengthen previous identified loci and suggest new disease loci or pathways.

It is well known RA is caused by interactions between multiple genetic factors and environmental factors, and that a complex molecular network and different cellular pathways play key roles in development of RA (15). If a specific pathway is relevant to disease susceptibility, association signals would

be expected to be overrepresented for the SNPs in genes in the pathway (16). Given the limited power of GWAS to detect single SNP associations, we adopted a pathway-based approach. Pathway-driven approaches are more powerful, as they take into account biological interplay among genes, and are attractive as they provide insight as to how multiple genes might contribute to the pathogeneses of diseases (17).

In the present study, we identified 49 candidate SNPs included in 37 candidate pathways associated with RA. Furthermore, the top 5 candidate SNPs were all at HLA loci. These candidate SNPs and pathways indicated 22 hypothetical biological mechanisms. In this genome-wide search for pathways associated with RA, the most strongly associated pathway related to antigen processing and presentation of peptide antigen via HLA class II. This result was consistent with the well-known role of HLA in the pathogenesis of RA (15). The most significant SNP → gene → hypothesis was as follows: rs1063478 → HLA-DMA → antigen processing and presentation of peptide antigen via HLA class II, regulation of lymphocyte differentiation and activation, and adaptive immune response. When we excluded the HLA region, ICSNPathway analysis identified two candidate SNPs identified in ten candidate pathways, which provided two hypothetical biological mechanisms. First, rs2476601 alters PTPN22 in the context of the pathway of immune response-activation cell surface receptor signalling pathway, lymphocyte differentiation, protein tyrosine phosphatase activity, protein amino acid dephosphorylation, phosphoprotein phosphatase activity. Second, rs2230926 alters TNFAIP3 in the context of the pathway of regulation of I kappaB kinase NF kappaB cascase, and CD40L signalling pathway.

It is well known that PTPN22 and TNFAIP3 play key roles in susceptibility to RA (18, 19). The two hypotheses derived by GWAS data interpretation using ICSNPathway analysis mentioned above are well supported by experimental evidence. The 1858C→T SNP of PTPN22 (rs2476601) is one of the best examples of a non-HLA common susceptibility allele for autoimmunity (20, 21). The PTPN22 gene maps to chromosome 1p13.3-p13.1 and encodes a lymphoid-specific phosphatase (Lyp). Lyp is an intracellular PTP, and physically binds via its proline-rich motif to the SH3 domain of Csk kinase, which is an important suppressor of kinases that mediate T cell activation (22). The PTPN22 1858C->T SNP changes the amino acid at position 620 from an arginine (R) to a tryptophan (W), disrupts the interaction between Lyp and Csk, and thus, inhibits complex formation and suppresses T cell activation. *In vitro* experiments have shown that the T allele of PTPN22 binds less efficiently to Csk than the C allele, which suggests that T cells expressing the T allele may be hyperresponsive, and that consequently, individuals carrying this allele may be prone to autoimmunity (23, 24). Actually, a meta-analysis conducted by Lee *et al*. (25) showed that the PTPN22 C1858T polymorphism is associated with RA susceptibility in Europeans.

TNFAIP3 encodes ubiquitin-editing protein A20, which is an inhibitor of nuclear factor kappa B (NF-κB) activity in several signalling pathways, including those of TNF and Toll-like receptors (26). Furthermore, A20-deficient mice were found to show systematic inflammation, damage involving kidneys and joints, and to develop autoimmunity (27). TNFAIP3 participates in the negative regulation of inflammatory responses, and alterations in the activity or expression of TNFAIP3-encoded A20 may influence the pathogenesis of RA (28). The TNFAIP3 gene (located at 6q23) is known to be associated with susceptibility to multiple autoimmune diseases (29). In particular, rs2230926 is located in the coding region of TNFAIP3 and an amino acid substitution of Phe to Cys at position 127 in the ovarian tumour domain has been suggested to play a role in the inhibitory function of A20 (30). Furthermore, the Cys127 allele product has been reported to be modestly less effective at inhibiting NF-κB activation by TNF than the Phe127 allele product (31), and associations between TNFAIP3 polymorphisms and RA have been reported in different ethnic groups (19).

The original meta-analysis identified previous known SNPs associated with RA, including HLA loci, PTPN22, and TNFAIP3 (6). Our pathway analysis using the meta-analysis dataset confirmed HLA, PTPN22, and TNFAIP3 as candidate genes of RA. The individual SNPs identified in this pathway analysis were different from the SNPs observed in the meta-analysis, but genes in genuinely associated pathways were consistently associated in the meta-analysis results.

In order to solve the challenge presented by GWAS data interpretation, pathway-based approaches, such as, ICSNPathway analysis, were developed (8). However, there is no straightforward way of comparing various pathway analysis methods against each other. The incomplete annotation of the human genome is an important limitation of the pathway-based approach for GWAS analysis. ICSNPathway analysis is not intended to predict true causal SNPs and pathways due to limited understanding of their genetic basis in complex diseases (8). A proportion of human genes remain uncharacterised, and thus, these genes have not been mapped to predicted pathways. Pathway analysis for GWAS data is not well developed, and thus, results should be interpreted with caution. Additional studies are needed to confirm the causal SNPs and genes underlying the association of pathways with RA identified during the present study. Nevertheless, pathway-based approaches play a complementary role in the identification of novel genes that confer disease susceptibility. Thus, the results obtained in the present study using ICSNPathway analysis may result in the formulation of new hypotheses for additional validations.

Conclusion

Summarising, we examined the meta-analysis GWAS results of 6 RA GWAS datasets to identify genetic associations with RA at both the SNP and pathway levels. Pathway analysis indicated candidate SNPs and genes identified in the pathways involving HLA, PTPN22, and TNAIP3 associated with RA susceptibility. Further studies are needed to confirm and explore the genetic variations of molecular pathways in RA.

References

1. HARRIS ED JR.: Rheumatoid arthritis. Pathophysiology and implications for therapy. *N Engl J Med* 1990; 322: 1277-89.
2. KIM HR, PARK MK, CHO ML *et al*.: Induction of macrophage migration inhibitory factor in ConA-stimulated rheumatoid arthritis synovial fibroblasts through the P38 map kinase-dependent signaling pathway. *Korean J Intern Med* 2010; 25: 317-26.
3. MACGREGOR AJ, SNIEDER H, RIGBY AS *et al*.: Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum* 2000; 43: 30-7.
4. MANOLIO TA: Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010; 363: 166-76.
5. JOHNSON AD, O'DONNELL CJ: An open access database of genome-wide association results. *BMC Med Genet* 2009; 10: 6.
6. STAHL EA, RAYCHAUDHURI S, REMMERS EF *et al*.: Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* 2010; 42: 508-14.
7. WANG K, LI M, HAKONARSON H: Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* 2010; 11: 843-54.
8. ZHANG K, CHANG S, CUI S, GUO L, ZHANG L, WANG J: ICSNPathway: identify candidate causal SNPs and pathways from genome-wide association study by one analytical framework. *Nucleic Acids Res* 2011; 39: W437-43.
9. ZEGGINI E, IOANNIDIS JP: Meta-analysis in genome-wide association studies. *Pharmacogenomics* 2009; 10: 191-201.
10. REINER A, YEKUTIELI D, BENJAMINI Y: Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 2003; 19: 368-75.
11. KANEHISA M, GOTO S, FURUMICHI M, TANABE M, HIRAKAWA M: KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 2010; 38: D355-60.
12. ASHBURNER M, BALL CA, BLAKE JA *et al*.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; 25: 25-9.
13. JOHNSON AD, HANDSAKER RE, PULIT SL, NIZZARI MM, O'DONNELL CJ, DE BAKKER PI: SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 2008; 24: 2938-9.
14. ELBERS CC, VAN EIJK KR, FRANKE L *et al*.: Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet Epidemiol* 2009; 33: 419-31.
15. CHOI SJ, RHO YH, JI JD, SONG GG, LEE YH: Genome scan meta-analysis of rheumatoid arthritis. *Rheumatology* (Oxford) 2006; 45: 166-70.
16. FRIDLEY BL, BIERNACKA JM: Gene set analysis of SNP data: benefits, challenges, and future directions. *Eur J Hum Genet* 2011; 19: 837-43.
17. PEDROSO I, BREEN G: Gene set analysis and network analysis for genome-wide association studies. Cold Spring Harb Protoc 2011; 2011:pdb.top065581.
18. LEE YH, RHO YH, CHOI SJ *et al*.: The PTPN22 C1858T functional polymorphism and autoimmune diseases--a meta-analysis. *Rheumatology* (Oxford) 2007; 46: 49-56.
19. HUGHES LB, REYNOLDS RJ, BROWN EE *et al*.: Most common single-nucleotide polymorphisms associated with rheumatoid arthritis in persons of European ancestry confer risk of rheumatoid arthritis in African Americans. *Arthritis Rheum* 2010; 62: 3547-53.
20. SIMINOVITCH KA: PTPN22 and autoimmune disease. *Nat Genet* 2004; 36: 1248-9.
21. GREGERSEN PK: Pathways to gene identification in rheumatoid arthritis: PTPN22 and beyond. *Immunol Rev* 2005; 204: 74-86.
22. COHEN S, DADI H, SHAOUL E, SHARFE N, ROIFMAN CM: Cloning and characterization of a lymphoid-specific, inducible human protein tyrosine phosphatase, Lyp. *Blood* 1999; 93: 2013-24.
23. BOTTINI N, MUSUMECI L, ALONSO A *et al*.: A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat Genet* 2004; 36: 337-8.
24. BEGOVICH AB, CARLTON VE, HONIGBERG LA *et al*.: A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet* 2004; 75: 330-7.
25. LEE YH, BAE SC, CHOI SJ, JI JD, SONG GG: The association between the PTPN22 C1858T polymorphism and rheumatoid arthritis: a meta-analysis update. *Mol Biol Rep* 2012; 39: 3453-60.
26. BOONE DL, TURER EE, LEE EG *et al*.: The ubiquitin-modifying enzyme A20 is required for termination of Toll-like receptor responses. *Nat Immunol* 2004; 5: 1052-60.
27. LEE EG, BOONE DL, CHAI S *et al*.: Failure to regulate TNF-induced NF-kappaB and cell death responses in A20-deficient mice. *Science* 2000; 289: 2350-4.
28. TAVARES RM, TURER EE, LIU CL *et al*.: The ubiquitin modifying enzyme A20 restricts B cell survival and prevents autoimmunity. *Immunity* 2010; 33: 181-91.
29. DIEUDE P, GUEDJ M, WIPFF J *et al*.: Association of the TNFAIP3 rs5029939 variant with systemic sclerosis in the European Caucasian population. *Ann Rheum Dis* 2010; 69: 1958-64.
30. COORNAERT B, CARPENTIER I, BEYAERT R: A20: central gatekeeper in inflammation and immunity. *J Biol Chem* 2009; 284: 8217-21.
31. MUSONE SL, TAYLOR KE, LU TT *et al*.: Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus. *Nat Genet* 2008; 40: 1062-4.