A Systems Genetics Approach Provides a Bridge from Discovered Genetic Variants to Biological Pathways in Rheumatoid Arthritis

Hirofumi Nakaoka[1], Tailin Cui[2], Atsushi Tajima[2,3], Akira Oka[2], Shigeki Mitsunaga[2], Koichi Kashiwase[4], Yasuhiko Homma[5], Shinji Sato[6], Yasuo Suzuki[6], Hidetoshi Inoko[2], Ituro Inoue[1,2]*

1 Division of Human Genetics, Department of Integrated Genetics, National Institute of Genetics, Mishima, Shizuoka, Japan, **2** Division of Molecular Life Science, School of Medicine, Tokai University, Isehara, Kanagawa, Japan, **3** Department of Human Genetics and Public Health, Institute of Health Biosciences, The University of Tokusima Graduate School, Tokushima, Tokushima, Japan, **4** Department of Laboratory, Japanese Red Cross Tokyo Blood Center, Koto-ku, Tokyo, Japan, **5** Department of Clinical Health Science, Tokai University School of Medicine, Isehara, Kanagawa, Japan, **6** Department of Internal Medicine, Division of Rheumatology, Tokai University School of Medicine, Isehara, Kanagawa, Japan

Abstract

Genome-wide association studies (GWAS) have yielded novel genetic loci underlying common diseases. We propose a systems genetics approach to utilize these discoveries for better understanding of the genetic architecture of rheumatoid arthritis (RA). Current evidence of genetic associations with RA was sought through PubMed and the NHGRI GWAS catalog. The associations of 15 single nucleotide polymorphisms and *HLA-DRB1* alleles were confirmed in 1,287 cases and 1,500 controls of Japanese subjects. Among these, *HLA-DRB1* alleles and eight SNPs showed significant associations and all but one of the variants had the same direction of effect as identified in the previous studies, indicating that the genetic risk factors underlying RA are shared across populations. By receiver operating characteristic curve analysis, the area under the curve (AUC) for the genetic risk score based on the selected variants was 68.4%. For seropositive RA patients only, the AUC improved to 70.9%, indicating good but suboptimal predictive ability. A simulation study shows that more than 200 additional loci with similar effect size as recent GWAS findings or 20 rare variants with intermediate effects are needed to achieve AUC = 80.0%. We performed the random walk with restart (RWR) algorithm to prioritize genes for future mapping studies. The performance of the algorithm was confirmed by leave-one-out cross-validation. The RWR algorithm pointed to *ZAP70* in the first rank, in which mutation causes RA-like autoimmune arthritis in mice. By applying the hierarchical clustering method to a subnetwork comprising RA-associated genes and top-ranked genes by the RWR, we found three functional modules relevant to RA etiology: ''leukocyte activation and differentiation'', ''pattern-recognition receptor signaling pathway'', and ''chemokines and their receptors''. These results suggest that the systems genetics approach is useful to find directions of future mapping strategies to illuminate biological pathways.

Citation: Nakaoka H, Cui T, Tajima A, Oka A, Mitsunaga S, et al. (2011) A Systems Genetics Approach Provides a Bridge from Discovered Genetic Variants to Biological Pathways in Rheumatoid Arthritis. PLoS ONE 6(9): e25389. doi:10.1371/journal.pone.0025389

Editor: Amr H. Sawalha, University of Oklahoma and Oklahoma Medical Research Foundation, United States of America

Received July 11, 2011; **Accepted** September 2, 2011; **Published** September 28, 2011

Copyright: © 2011 Nakaoka et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported in part by a grant-in-aid for scientific research from the Japanese Ministry of Health, Labor, and Welfare. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. No additional external funding was received for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: itinoue@lab.nig.ac.jp

Introduction

Genome-wide association studies (GWAS) have identified a large number of novel genetic loci underlying susceptibility to common diseases [1], which leads to an interest in how these discoveries may be translated into improvement in health care and public health. Identification of associated variants can illuminate causal pathways and provide a clue for therapeutic targets [2]. Ultimately, it may be possible to predict the development of common diseases by genetic profiling, in which multiple genetic loci are simultaneously tested [3].

There are conflicting views regarding the usefulness of genetic variants in disease prediction [4–8]. The idea widely received is that the predictive ability of genetic profiling is limited with some exceptions [5] because most common genetic variants identified to date confer relatively small effects on disease risk and explain a small portion of the individual variation in disease risks [9]. The risk estimates will be updated and become more accurate with new genetic discoveries by conducting more large-scale GWAS [6] and by extending the analysis of low frequency and rare variants [10]. There are some examples that individually rare variants with relatively large effect contribute to complex trait variation [11–13]. It is important to infer the allelic architecture of as-yet-discovered risk variants on the basis of current evidence of known disease-associated variants in order to provide clues for future mapping strategies [14].

There are prerequisites for evidence-based genetic testing. First, a rigorous scientific basis for the genetic variants used for the genetic profiling is essential [15]. In fact, most of the genetic variants used by direct-to-consumer genetic testing to predict an individual's risk to common diseases have been shown to lack consistent evidence of gene-disease associations [15]. Second, and probably most importantly, the predictive ability of genetic variants should be evaluated [5]. The predictive ability can be quantified by several measures such as the area under the receiver operating characteristic curve [16]. Third, it is necessary to corroborate the generalizability of a genetic risk prediction model in independent datasets [17]. Systematic validation and characterization of the evidence of genetic associations at both discovery and translational phases of human genomics are also required [18,19]. In these circumstances, meta-analysis can be a useful tool to improve the estimation of effect sizes of genetic variants by combining results from individual studies, thereby making it possible to evaluate variants for model inclusion in a rigorous way [20].

We propose here a systems genetics approach to utilize current evidence of genetic associations for better understanding of the genetic architecture of complex disease [21]. The outline of our approach is schematically shown in Figure 1 (The left and right columns correspond to the first three and last steps in the following description). First, genetic variants associated with the disease of interest are identified by exhaustively reviewing meta-analyses of genetic association studies. Second, the association and the predictive ability of the selected variants are confirmed in real case-control subjects. Third, a framework of simulation study is formulated to address how many additional loci should be mapped for the establishment of acceptable levels of genetic risk prediction. Fourth, a network analysis is implemented where information on disease-associated genes are integrated through human interactome such as the protein-protein interaction (PPI) network for the design of future mapping studies and exploring biological pathways [22].

We applied the systems genetics approach to rheumatoid arthritis (RA, [MIM 180300]). RA is a common autoimmune disease characterized by chronic, destructive and debilitating arthritis [23]. The etiology of RA is not completely known and most likely involves a complex interplay of both genetic and environmental factors. It has been shown that multiple alleles at the *HLA-DRB1* locus within the major histocompatibility complex (MHC) region are associated with RA. RA susceptibility loci outside the MHC region have been identified through candidate gene approaches and GWAS [24,25]. The subdivision of RA patients in terms of the presence or absence of rheumatoid factor (RF) and antibodies against cyclic citrullinated peptide (anti-CCP) is increasingly recognized for possible prevention and treatment strategies. Genetic factors may also contribute to the phenotypic diversity in RA [26].

Results

Electronic database searches

We sought published meta-analyses that had evaluated the association between genetic variants and RA risk in population-based studies through two electronic databases: PubMed and NHGRI GWAS catalog. Figure S1A shows the outline of our literature search strategy using PubMed database. The reasoning for each of the excluded articles in the abstract reading, full-text search and data extraction stage is listed in Tables S1, S2, and S3, respectively. After selecting meta-analyses that fulfilled inclusion criteria, we found 29 articles addressing 27 variants located on 18 genetic loci [27–55]. After reducing redundant variants on the same genetic locus, 20 variants were identified (Text S1A). We also retrieved seven articles addressing the contribution of the *HLA-DRB1* locus [56–62].

In order to overview the retrieved meta-analyses, we classified individual studies analyzing the same genetic variants into three groups: studies showing significant evidence of increased and reduced risk, or non-significant result (Figure 2). In cases of single nucleotide polymorphism (SNP) rs7574865 at the *STAT4* locus, rs2476601 at the *PTPN22* locus, and rs6920220 and rs10499194 at the *TNFAIP3-OLIG3* locus, consistent lines of evidence of associations were observed. In the other cases, more than half of the individual studies did not show significant evidence of association. However, the direction of associations in the studies showing significant evidence was consistent for each variant except for rs1800629 at *TNF-α* and rs396991 at *FCGR3A*. This result suggests that most of the individual studies may be underpowered to detect small genetic effect [63,64]. Thus, conclusions derived from meta-analyses may be useful to select genetic variants for risk prediction models.

The outline of the NHGRI GWAS catalog search is shown in Figure S1B. Eight articles were retrieved [52,65–71]. We found the 61 associations with $P<1.0\times10^{-5}$: 7 for the HLA region and 54 for the non-HLA region comprising 34 distinct genetic loci. Restricting the statistical significance level at $P<5.0\times10^{-8}$, 18 associations, 10 of which did not overlap those from the PubMed search, were retrieved. All of the retrieved associations were derived from meta-analyses of several GWAS and replication studies [69–71].

Re-analysis of published meta-analyses and selection of genetic variants

We re-analyzed the meta-analyses addressing 20 genetic associations (Table S4; Text S1A). For each meta-analysis, a median of 6,758 cases (interquartile range [IQR]: 3,445–10,994) and 7,643 (IQR: 3,367–14,406) controls had been involved. We found that there were 10 meta-analyses showing statistically significant between-study heterogeneity. This indicates that the between-study heterogeneity was more frequent than what would be expected by chance ($P=7.2\times10^{-6}$). The median of I^2 metric was 40.4% (IQR: 16.1–63.9%). In 14 of 20 meta-analyses, the genetic associations passed the significance threshold of $P=2.5\times10^{-3}$ under the fixed effects model meta-analysis in the overall populations. Even when applying the random effects model meta-analysis, which is a conservative approach under the presence of between-study heterogeneity, evidence of association was confirmed in all of the 14 polymorphisms ($P<0.05$).

From the PubMed search, we identified the following 14 variants that fulfilled our selection criteria: rs7574865 (*STAT4*); rs3087243 (*CTLA4*); rs7528684 (*FCRL3*); rs3761847 (*TRAF1-C5*); rs2812378 (*CCL21*); rs4810485 (*CD40*); rs42041 (*CDK6*); rs2240340 (*PADI4*); rs2476601 (*PTPN22*); rs2073838 (*SLC22A4*); rs2004640 (*IRF5*); rs6920220 and rs10499194 (*TNFAIP3-OLIG3*); and rs333 (*CCR5*). Among these 14 polymorphisms, 13 were SNPs and one was the 32 bp-deletion polymorphism in *CCR5* (referred to as rs333).

For the *HLA-DRB1* alleles, we selected six alleles that were significantly associated with RA risk in a comprehensive review article [62] using the largest collection of relevant articles: *HLA-DRB1*01:01*, *DRB1*09:01*, *DRB1*10:01*, *DRB1*04:04*, *DRB1*04:01*, and *DRB1*04:05*.

We identified an additional 10 SNPs from the NHGRI GWAS catalog: rs3093024 (*CCR6*); rs874040 (*RBPJ*); rs11676922 (*AFF3*); rs13017599 (*REL*); rs6859219 (*ANKRD55*); rs934734 (*SPRED2*); rs2736340 (*BLK*); rs26232 (*C5orf30*); rs13315591 (*FAM107A*); and rs706778 (*IL2RA*).

Collectively, 23 SNPs, one deletion polymorphism, and six *HLA-DRB1* alleles that were significantly associated with RA risk were identified. Allele frequencies of the genetic variants with validated associations with RA were highly differentiated between

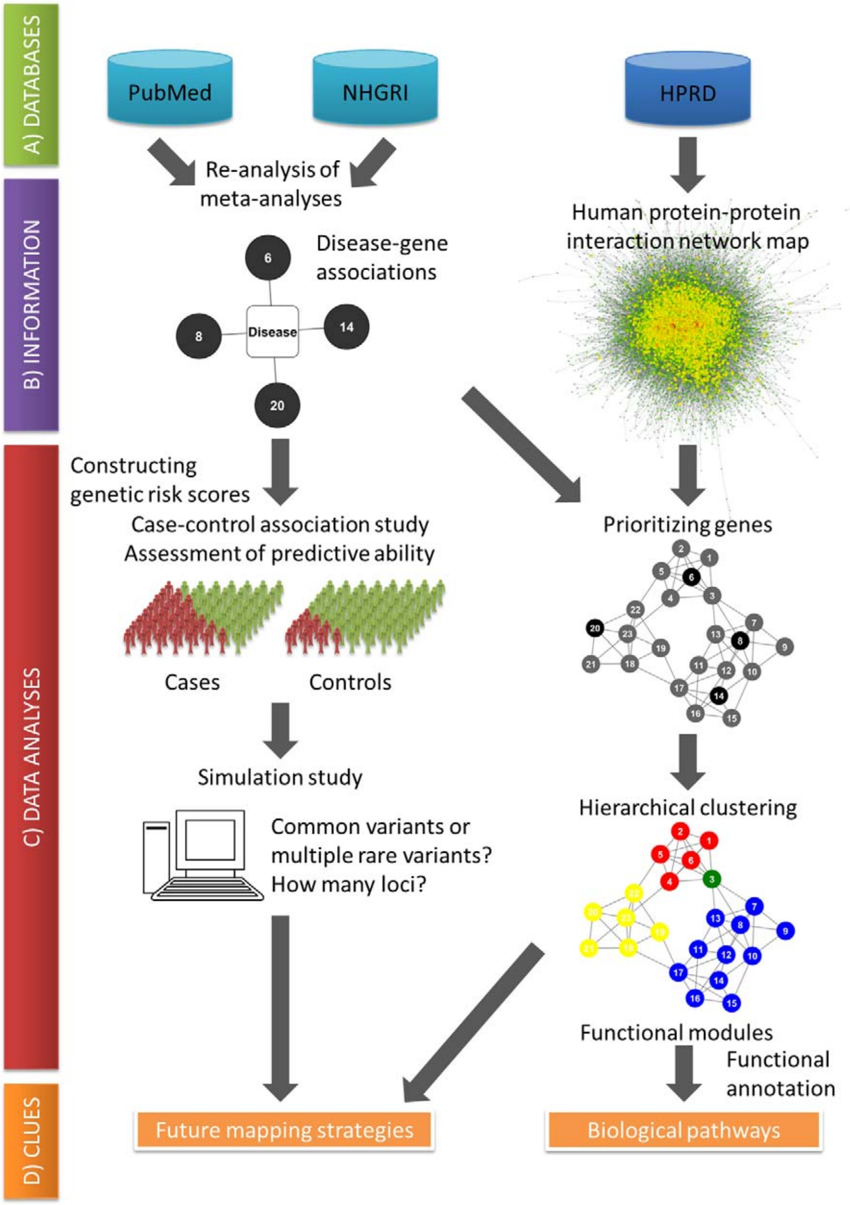
A) DATABASES
B) INFORMATION
C) DATA ANALYSES
D) CLUES
PubMed
NHGRI
HPRD
Re-analysis of meta-analyses
Disease-gene associations
Disease
6
8
14
20
Human protein-protein interaction network map
Constructing genetic risk scores
Case-control association study
Assessment of predictive ability
Cases
Controls
Simulation study
Common variants or multiple rare variants? How many loci?
Prioritizing genes
Hierarchical clustering
Functional modules
Functional annotation
Future mapping strategies
Biological pathways

Figure 1. The systems genetics approach proposed in this study. A) Databases from which knowledge is extracted. Meta-analyses and GWAS findings are sought in PubMed and NHGRI GWAS catalog, respectively. Human protein-protein interaction data is obtained from HPRD. B) Retrieved information is used to create two types of networks: 'gene-disease association network' and 'protein-protein interaction network'. C) The data analysis phase. The gene-disease associations are confirmed by using real case-control subjects. The predictive ability of selected genetic variants is evaluated and the result is used in the simulation study to infer allelic architecture of as-yet-discovered genetic variants. Two types of networks are integrated to prioritize genes by the global measure of distance to known disease-associated genes within the protein-protein interaction network. Hierarchical clustering algorithm is applied to a subnetwork comprising top-ranked genes and functional annotation for each cluster is used for the inference on biological pathways underlying the disease of interest. D) The systems genetics approach emerges two types of clues: Future mapping strategies, and biological pathways.
doi:10.1371/journal.pone.0025389.g001

East Asian and European populations (Table S5). Among them, 15 SNPs with minor allele frequency greater than 5% in Japanese and six *HLA-DRB1* alleles were selected through our database searches.

Ethnic differences

We examined the ethnicity-specific effects of these variants (Table S6). We found heterogeneity in the odds ratios (ORs) between ethnic groups at $P<0.05$ for rs2240340 (*PADI4*) and rs7528684 (*FCRL3*). The OR of rs2240340 was larger for East Asian (OR = 1.31, 95% confidence interval [CI]; 1.22–1.41, $P = 5.6\times10^{-13}$) than for European descent populations (OR = 1.03, 95% CI; 0.99–1.07, $P = 0.16$). Similarly, the rs7528684 association was stronger for East Asian populations (OR = 1.16, 95% CI; 1.09–1.24, $P = 7.8\times10^{-6}$) than for European descent populations (OR = 1.03, 95% CI; 0.98–1.09, $P = 0.27$). The effects observed with East Asian populations were used in the genetic risk score (Table 1).

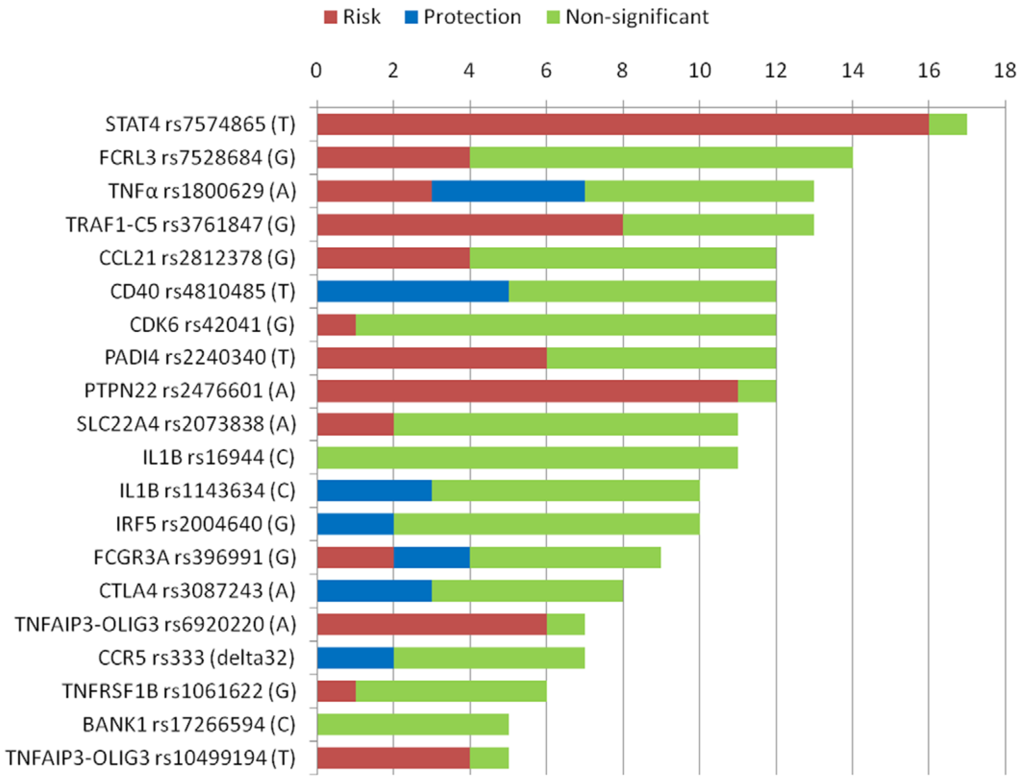

Figure 2. Overview of association studies in RA of 20 genetic variants examined in the meta-analyses met our inclusion criteria. Colored bars displays number of individual studies according to the result of testing for association of each variant with RA: red, studies show significant evidence of an increased risk; blue, studies show significant evidence of disease protection; and green, studies show non-significant result. The significance level was set at $P = 0.05$.
doi:10.1371/journal.pone.0025389.g002

Association analysis

We conducted a case-control study of 1,287 RA cases and 1,500 controls in Japanese (see Materials and Methods for description of our cohort). Genotype counts for six *HLA-DRB1* alleles and 15 SNPs are shown in Table S7. For the SNPs, the missing genotype rates were small (at most 1.3% for rs2736340). SNP rs2004640 (*IRF5*), which deviated from the HWE in controls at $P<0.001$, was excluded from subsequent analyses.

We assessed the association of each genetic variant with RA risk by logistic regression analysis (Table 1). For the *HLA-DRB1* alleles, *HLA-DRB1*04:05* allele showed highly significant evidence of association with the risk of RA. It should be noted that the ORs of all the *HLA-DRB1* alleles in the multivariate logistic regression analysis were larger than those in the univariate analysis. When an allele was evaluated in the univariate analysis, the other five putative risk alleles were grouped together into one referent group, which resulted in a weakened association signal.

For the SNPs, strong evidence of association was observed with rs3093024 in *CCR6* ($P = 4.1\times10^{-5}$, OR = 1.25), rs2240340 in *PADI4* ($P = 1.5\times10^{-4}$, OR = 1.23), rs2736340 in *BLK* ($P = 3.2\times10^{-4}$, OR = 1.24), and rs4810485 in *CD40* ($P = 4.7\times10^{-4}$, OR = 0.80). We found that four SNPs showed nominally significant associations at $P<0.05$ for rs26232 (*C5orf30*), rs2073838 (*SLC22A4*), rs11676922 (*AFF3*), and rs7528684 (*FCRL3*). SNPs on *SPRED2* and *STAT4* showed suggestive associations at $P<0.1$. SNPs on *CTLA4*, *TRAF1-C5*, and *IL2RA* had the same direction of effect as identified in previous studies. SNP rs10499194 on *TNFAIP3-OLIG3* showed the opposite direction of effect. The observed opposite direction of rs10499194 seems to be attributable to the difference in linkage disequilibrium between marker and true disease allele across populations. Shimane et al. showed a similar result and identified a non-synonymous SNP (rs2230926) in *TNFAIP3* associated with RA [72].

The ORs for SNPs obtained with the univariate analysis were similar to those with the multivariate analysis, indicating that the associations of these SNPs are independent association signals. These results suggest that a substantial proportion of the loci identified in the meta-analyses are likely to be shared across populations.

Discrimination using genetic risk models

This study is reported in accordance with the Strengthening the Reporting of Genetic Risk Prediction Studies recommendations [73]. With the use of the receiver operating characteristic (ROC) curve, we calculated the area under the ROC curve (AUC) to evaluate the predictive ability of the genetic risk scores based on the selected variants (see Materials and Method for description of the construction of genetic risk score). The AUC for the HLA model was 65.9% (95% CI, 63.9 to 67.9%). The non-HLA model including 14 SNPs showed an AUC of 58.8% (56.6 to 60.9%). The AUC for the integrative model was 68.4% (66.4 to 70.4%). The addition of 14 SNPs to the *HLA-DRB1* alleles increased the AUC by 2.5%. The observed increase in the AUC was statistically significant ($P = 2.8\times10^{-6}$). The integrative model shows better fit than the HLA model in terms of Akaike's information criterion (Table 2). We examined an *ad hoc* model, where rs10499194 on

Table 1. Association analysis of RA with selected genetic variants.

Gene	SNP	A1/A2[A]	Univariate[B]		Multivariate[B]		Previous report[C]
			OR (95% CI)	***P***	**OR (95% CI)**	***P***	**OR (95% CI)**
HLA-DRB1	**01:01*	+/−	1.29 (1.03–1.61)	0.025	1.95 (1.52–2.48)	8.8×10^{-8}	1.60 (1.39–1.84)
	**09:01*	+/−	1.20 (1.04–1.39)	0.012	1.74 (1.48–2.04)	1.8×10^{-11}	1.67 (1.44–1.94)
	**10:01*	+/−	2.88 (1.42–5.83)	3.3×10^{-3}	3.59 (1.72–7.52)	7.0×10^{-4}	2.35 (1.90–2.91)
	**04:01*	+/−	1.89 (1.23–2.90)	3.9×10^{-3}	2.70 (1.69–4.30)	3.0×10^{-5}	3.30 (3.01–3.61)
	**04:04*	+/−	1.49 (0.55–4.02)	0.43	2.92 (0.98–8.67)	0.054	1.85 (1.54–2.22)
	**04:05*	+/−	2.31 (2.01–2.66)	1.3×10^{-31}	2.80 (2.40–3.27)	9.4×10^{-39}	3.84 (3.30–4.46)
SNPs with strong evidence of association (*P*<2.5×10^{-3})							
CCR6	rs3093024	A/G	1.25 (1.12–1.39)	4.1×10^{-5}	1.26 (1.13–1.42)	6.3×10^{-5}	1.19 (1.15–1.24)
PADI4	rs2240340	T/C	1.23 (1.11–1.37)	1.5×10^{-4}	1.24 (1.11–1.40)	2.6×10^{-4}	1.31 (1.22–1.41)
BLK	rs2736340	T/C	1.24 (1.10–1.39)	3.2×10^{-4}	1.24 (1.09–1.41)	7.8×10^{-4}	1.19 (1.13–1.27)
CD40	rs4810485	T/G	0.80 (0.72–0.89)	4.7×10^{-4}	0.82 (0.73–0.92)	7.8×10^{-4}	0.87 (0.83–0.90)
SNPs with nominally significant association signals (*P*<0.05)							
C5orf30	rs26232	T/C	0.86 (0.77–0.98)	0.018	0.86 (0.75–0.98)	0.021	0.90 (0.87–0.94)
SLC22A4	rs2073838	A/G	1.14 (1.02–1.27)	0.022	1.17 (1.04–1.32)	0.012	1.11 (1.05–1.18)
AFF3	rs11676922	T/A	1.11 (1.00–1.24)	0.043	1.11 (0.99–1.24)	0.083	1.14 (1.10–1.18)
FCRL3	rs7528684	G/A	1.11 (1.00–1.24)	0.047	1.08 (0.96–1.21)	0.20	1.16 (1.09–1.24)
SNPs showing the same direction of effect							
SPRED2	rs934734	G/A	1.14 (0.99–1.31)	0.064	1.17 (1.00–1.36)	0.043	1.13 (1.09–1.17)
STAT4	rs7574865	T/G	1.10 (0.98–1.23)	0.093	1.09 (0.97–1.23)	0.14	1.23 (1.19–1.27)
CTLA4	rs3087243	A/G	0.92 (0.82–1.04)	0.18	0.96 (0.84–1.10)	0.56	0.89 (0.85–0.95)
TRAF1	rs3761847	A/G	1.05 (0.95–1.17)	0.35	1.03 (0.91–1.15)	0.66	1.13 (1.09–1.17)
IL2RA	rs706778	T/C	1.05 (0.95–1.17)	0.36	1.05 (0.93–1.17)	0.43	1.12 (1.09–1.16)
SNPs showing the opposite direction of effect							
TNFAIP3	rs10499194	T/C	1.18 (0.96–1.46)	0.11	1.18 (0.94–1.48)	0.15	0.82 (0.77–0.87)

[A] A1 and A2 represent the coded and non-coded alleles, respectively.

[B] ORs and 95% CIs were estimated by logistic regression analyses using univariate analysis for each allele and then using multivariate analysis including all the alleles. The number of coded alleles (A1) was used as the predictor value in the logistic regression analyses.

[C] ORs and 95% CIs were calculated by meta-analyses of published studies: *HLA-DRB1* from [62]; *CD40*, *SLC22A4*, *STAT4*, *CTLA4*, *TRAF1*, *TNFAIP3*, and *IRF5* from re-analysis of meta-analyses shown in Table S4; *PADI4* and *FCRL3* from re-analysis of ethnicity-specific meta-analyses shown in Table S6; and *CCR6*, *BLK*, *C5orf30*, *AFF3*, *SPRED2*, and *IL2RA* from original GWASs [69–71]. These ORs were used to create genetic risk scores.

doi:10.1371/journal.pone.0025389.t001

TNFAIP3-OLIG3 showing the opposite effect as identified in previous studies was removed. The AUC was then 68.6 (66.6 to 70.6%). The improvement in the AUC from the integrative model was statistically significant ($P = 0.034$).

We performed the same ROC analyses by using only the patients with both anti-CCP and RF positivity (Table 2). The AUC for the HLA, non-HLA and integrative models was 68.3% (65.2 to 71.4%), 60.0% (56.7 to 63.3%), and 70.9% (67.8 to 73.9%), respectively. For each genetic risk model, the AUC in both RF and anti-CCP positive patients versus controls was greater than that in overall patients versus controls. The result of the association study for anti-CCP and RF positive RA is shown in Table S8.

Figure 3 depicts the distribution of genetic risk scores by phenotypic status for the integrative model. The distribution of the genetic risk scores in cases differs from that in controls ($P = 1.1\times10^{-61}$). The curve in RF and anti-CCP positive cases shifts upward compared to the curve in overall cases, indicating that the risk scores in RF and anti-CCP positive cases were larger than those in overall cases. This was reflected in better discrimination ability between both RF and anti-CCP positive patients and controls (AUC = 70.9%) than that between overall cases and controls (AUC = 68.4%). Each curve in Figure 3 looks like multimodal distribution. The multimodality of these curves is attributable to differences in genetic risk score among the *HLA-DRB1* genotypes.

Table 2. The discriminative ability and the global model fit of three predictive models according to subphenotype of case patients.

Case phenotype	Model	AUC (95% CI)	AIC[A]
Overall	HLA	65.9 (63.9–67.9)	3477.7
	Non-HLA	58.8 (56.6–60.9)	3630.7
	Integrative	68.4 (66.4–70.4)	3421.9
RF & anti-CCP positive	HLA	68.3 (65.2–71.4)	1603.7
	Non-HLA	60.0 (56.7–63.3)	1694.2
	Integrative	70.9 (67.8–73.9)	1578.0

[A] Akaike's information criterion.

doi:10.1371/journal.pone.0025389.t002

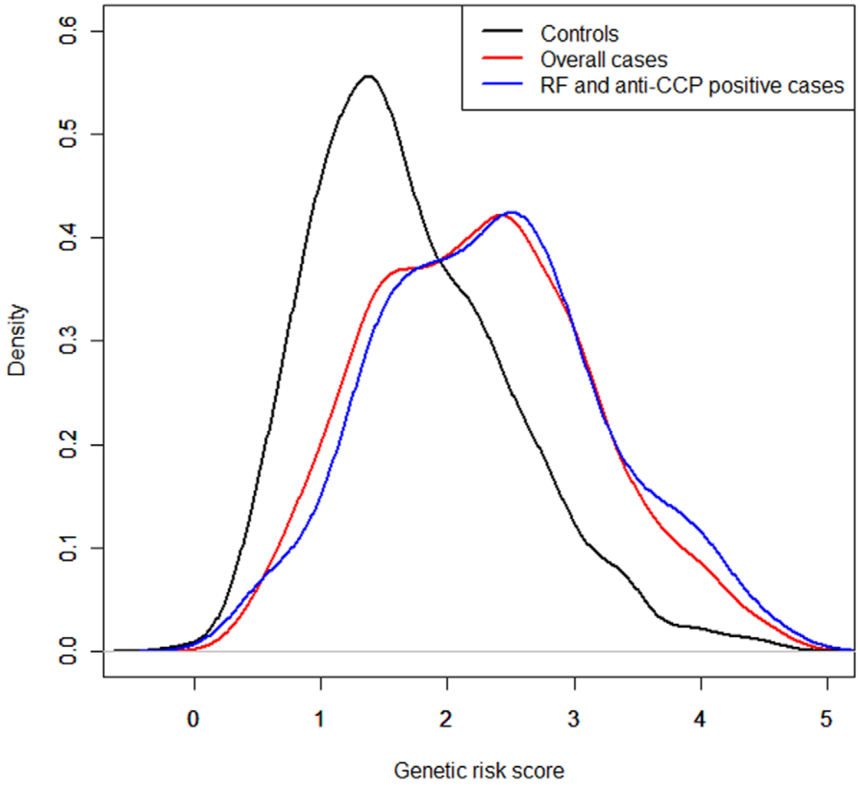

Figure 3. Distribution of risk scores by phenotypic status for the integrative model, in which six *HLA-DRB1* alleles and 14 SNPs were included. The curves were generated with a Gaussian kernel density smoother.
doi:10.1371/journal.pone.0025389.g003

Simulation study: How many additional loci should be mapped?

We investigated how many additional loci are required to achieve an acceptable level of genetic risk prediction via simulation study. We set AUC of 80.0% as an acceptable level based on the diagnostic accuracy of anti-CCP antibody and RF for RA. According to a recent meta-analysis [74], the pooled sensitivity and specificity were 67% and 95% for anti-CCP antibody, respectively, and 69% and 85% for IgM RF, respectively. The naïve estimate of the AUC was 81% for anti-CCP antibody and 75% for IgM RF.

We simulated the distribution of RA risks in the general population based on observed ORs and allele frequencies for the selected variants (see Materials and Methods, and Text S1B for details). For the base model in which 13,392,312 multi-locus genotypes generated by combining the six *HLA-DRB1* alleles and the 14 SNPs are included, the simulated AUC was 71.0%. The AUC of the base model was similar to that observed in anti-CCP and RF positive patients (AUC = 70.9%). Starting with the base model, we evaluated the simulated AUC value assuming that hypothetical additional loci were discovered.

Result of the simulation study is shown in Figure 4. Under the common disease-common variant hypothesis, ~50 loci are needed in the setting of additional loci with OR = 1.2 and risk allele frequency (RAF) of 0.30. Taking into consideration the fact that the ORs from recent GWAS of RA were close to 1.1, a scenario of OR = 1.1 and RAF = 0.30 may be more realistic. In this scenario, ~220 loci are required. When assuming the multiple rare variants with intermediate effects that remain undiscovered and setting the additional loci with OR of 3.0 and RAF of 0.01, only ~20 loci are sufficient for AUC of 0.80. When assuming OR = 2.0 and RAF = 0.01, an additional 50 loci are needed.

We further implemented simulations in which combination of common and rare variants was examined. When assuming *HLA-DRB1* alleles, 150 loci with OR = 1.1 and RAF = 0.30, and 10 loci with OR = 3.0 and RAF = 0.01, the AUC was 80.2%. The simulation rendered AUC = 95.2% under the assumption of *HLA-DRB1* alleles, 300 loci with OR = 1.1 and RAF = 0.30, and 140 loci with OR = 3.0 and RAF = 0.01.

Network analysis

The simulation study shows that many additional variants need to be discovered. We hypothesized that variants within genes on the same biological pathways of known RA susceptibility genes can be associated with RA. Then, we performed following network analyses to prioritize genes for future mapping studies.

We constructed the PPI network by using HPRD database [75,76]. The PPI network contained 37,080 interactions between 9,521 human proteins. The selected variants were assigned to a single protein-coding gene (Table S5; Text S1C). There are 19 RA-associated proteins mapped in the PPI network (HLA-DRB1, STAT4, FCRL3, TRAF1, CCL21, CD40, CDK6, PTPN22, SLC22A4, IRF5, CTLA4, TNFAIP3, CCR6, REL, SPRED2, BLK, FAM107A, and IL2RA).

We used the random walk with restart (RWR) algorithm [77] to prioritize genes in terms of the proximity to the validated RA susceptibility genes within the PPI network (see Materials and Method for details). As a preliminary test, we confirmed that the value of restart probability, *r*, did not largely affect the ranking of genes. When we examined different values of *r* (0.3, 0.5, and 0.7),

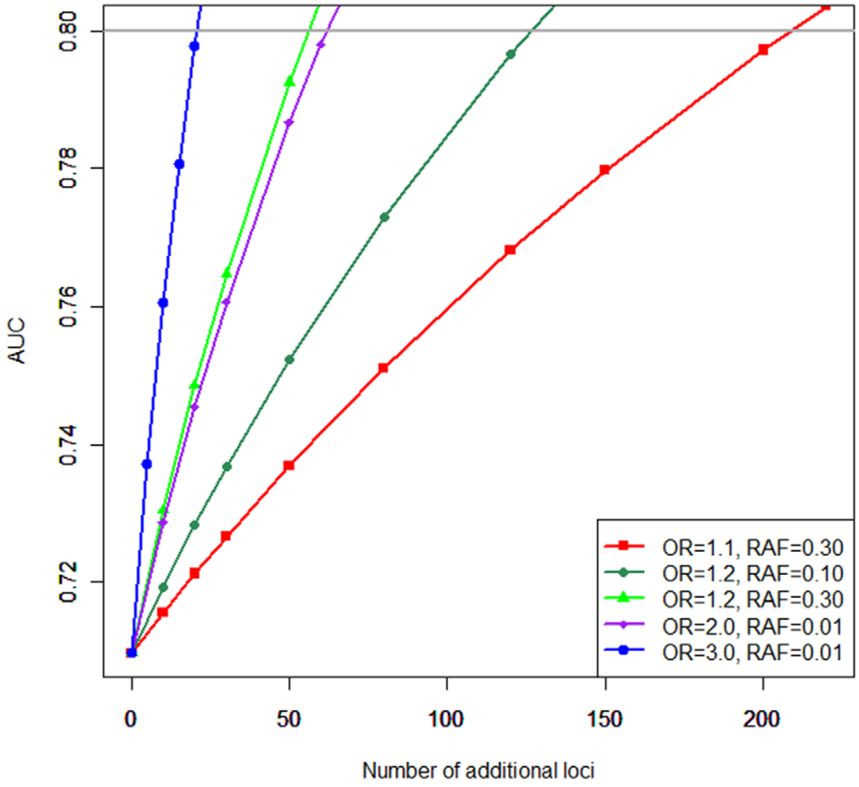

Figure 4. Simulation study addressing how many additional loci should be mapped for the establishment of excellent genetic risk prediction. Five scenarios with different combination of OR and RAF were examined.
doi:10.1371/journal.pone.0025389.g004

the spearman's rank correlation coefficients ranged from 0.967 to 0.993. The predictive ability of the network-guided gene prioritization method was evaluated by the leave-one-out cross-validation. As shown in Figure 5, most of the left-out genes are highly evaluated. For example, *TRAF1* ranked 58th among 9,503 genes evaluated. The AUC by the leave-one-out cross-validation was 84.4%, indicating an excellent predictive ability. This result also suggests that the RA-associated genes are in proximity to each other within the PPI network.

In the top-ranked genes, we can find many genes that may be involved in the susceptibility to RA and other autoimmune diseases. The RWR algorithm points to *ZAP70* in the first rank. Notably, a mutation in *ZAP70* is identified to cause chronic autoimmune arthritis in mice [78]. Sakaguchi et al. [78] demonstrate that the mutation in the mouse *ZAP70* affects thymic T-cell selection and leads to the development of RA-like arthritis. ZAP70 has direct interactions with PTPN22 and FCRL3 among proteins encoded by RA-associated genes in the HPRD database. FCRL3 has a direct interaction only with ZAP70 in the HPRD database, which might cause upward bias in the ranking of *ZAP70*. Even when excluding *FCRL3* from the list of seed vertices, *ZAP70* ranked 42th among 9,503 genes, indicating that the priority of the gene is robust and that ZAP70 is located proximal to proteins encoded by the RA-associated genes in the PPI network. We found that *CD247*, *IL2RB* and *IL2* ranked 32th, 34th and 39th, respectively, and have been associated with RA in follow-up study of GWAS [79,80] and studies exploring shared susceptibility loci among autoimmune diseases [81,82]. *CD80*, *FCGR2A*, *FCGR2B*, *ICAM1*, *JAK2*, *LYN*, *NFKBIA*, *PTPN11*, *STAT3* and *TRAF3IP2* were shown to be associated with other autoimmune diseases according to the NHGRI GWAS catalog and systematic review [83].

Figure 6A depicts an RA-associated network that is a subnetwork of the PPI network in which vertices are the RA-associated genes and genes ranked in the top 100 by the RWR algorithm and edges are physical interactions between their products. In order to detect functional modules in the RA-associated network, we applied the EAGLE algorithm [84] and found three complexes each containing more than 10 vertices (referred to as CL1-3). The CL1 and CL2 overlap each other.

We further explored functional annotations of these three clusters by using DAVID [85,86] (Table 3). The three clusters fitted into different categories of immunological pathway. CL1 can be assigned to an immunological pathway "leukocyte activation and differentiation" according to Gene Ontology (GO) terms annotated to genes in CL1 (Figure 6B). CL2 is associated with "pattern-recognition receptor signaling pathways" since genes in CL2 are enriched for GO terms and KEGG pathways such as Toll-like receptor and Nod-like receptor signaling pathways (Figure 6C). CL3 is enriched for genes relevant to "chemokines and their receptors" (Figure 6D). This result shows that the exploration of topology of the network based on curated disease susceptibility genes is useful to find functional modules involved in disease pathology. We confirmed that similar biological pathways are observed when the number of top ranked genes included into the RA-associated network is altered to 50 and 150 (Figures S2, S3, Tables S9, S10).

Discussion

The phenomenon named 'missing heritability' has received much attention [9] and calls into substantive question the usefulness of genetic profiles for disease risk prediction. In this

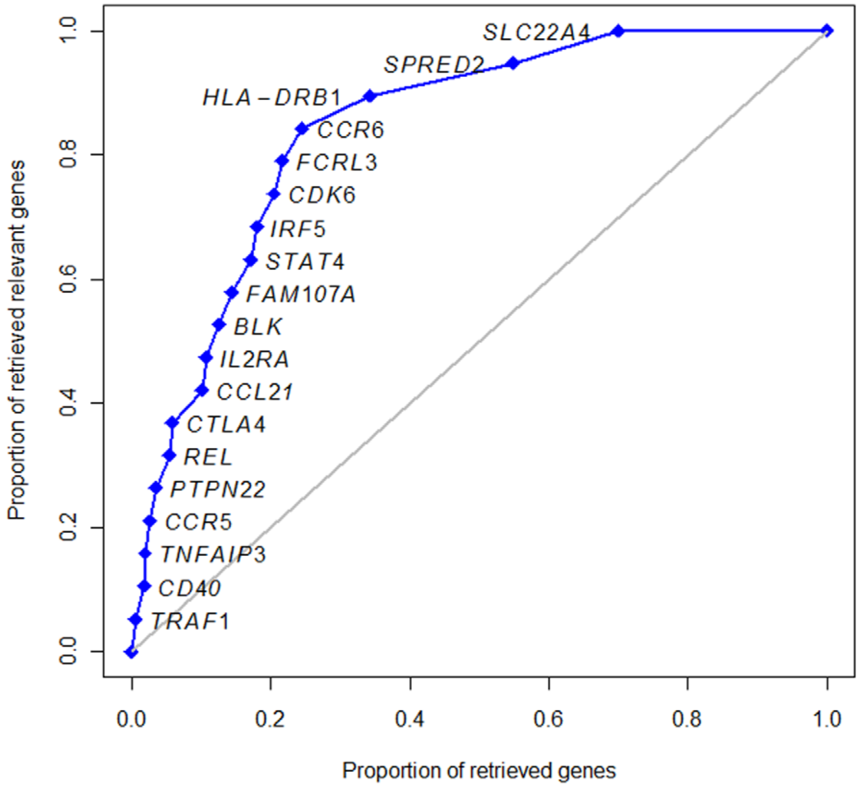

Figure 5. ROC curve using the leave-one-out cross-validation method to evaluate the predictive ability of the RWR algorithm. The gray diagonal line corresponds to the AUC of 0.5 and no discrimination (i.e., random performance).
doi:10.1371/journal.pone.0025389.g005

study, we performed a systematic approach to overview and validate current evidence of genetic associations with RA and utilized them to find directions of future mapping strategies.

One fundamental question is whether genetic risk factors for RA overlap across ethnic groups [87]. Although the associations of two SNPs (*PADI4* and *FCRL3*) were significantly stronger in East Asian than in European, these two SNPs represented significant association in the overall populations and the recent European GWA meta-analysis [70] captured weak association signals of these SNPs. This suggests that these SNPs may be common risk factors but that their contribution to RA risk may differ across ethnic groups. Furthermore, we confirmed that most of the selected genetic variants from meta-analyses and NHGRI GWAS catalog were consistently replicated in a case-control study of 1,287 RA cases and 1,500 controls of Japanese. These results suggest that a substantial proportion of the loci identified in the meta-analyses are likely to be shared across populations.

The predictive ability of genetic variants for the development of RA was moderate: the AUC for the integrative model was 68.4% (66.4 to 70.4%). Notably, the AUC improved to 70.9% (67.8 to 73.9%) when we used patients with both RF and anti-CCP positivity. This finding is consistent with European study (AUC = 71% [68 to 73%] for anti-CCP positive RA) although the list of selected variants used was slightly different [88]. This is the first study showing that the predictive ability of genetic variants for RA mainly derived from European GWAS is similar between European and non-European populations by using a substantial number of case and control subjects. However, the predictive ability is suboptimal at the current stage.

When we implemented a simulation study addressing how many additional loci should be mapped, we set a goal of genetic risk prediction that achieves a similar level of accuracy with anti-CCP antibody for RA. Such genetic risk prediction may have clinical utility: when patients have primary symptoms such as joint pain and stiffness, prior knowledge of their higher genetic risks for RA may inspire them to undergo highly specific diagnostic tests such as anti-CCP antibody. Early detection and treatment can prevent severe disability for many patients.

According to the simulation study, more than 200 loci with OR = 1.1 and RAF = 0.30 are required to achieve AUC of 80.0%, implying that efforts relying only on GWAS may be reaching limits for improving predictive ability. With the advent of the development of massively parallel DNA sequencing technologies, exploring rare variants of large effect has attracted increased attention [89]. There is some evidence of rare variants with a large impact on risk of RA and autoimmune diseases [13,90,91]. Functionally defective rare variants in *SIAE* were recently shown to be associated with autoimmune diseases including RA with ORs estimated at approximately 8.0 [91]. In our simulation study, additional 20 rare but not private variants (minor allele frequency of 1%) with intermediate effect (OR of 3.0) suffice for AUC of 80.0%. Several hundreds of cases and controls must be resequenced for the discovery of such rare variants. However, whole-genome and whole-exome sequencing of large samples are costly and otherwise infeasible. The use of bar-coded multiplexed and target enrichment sequencing of the exonic regions of hundreds of candidate genes [92,93] could be an alternative strategy if appropriate candidate genes were selected.

We applied the RWR algorithm to prioritize genes by using information on the list of curated RA-associated genes and the PPI network from HPRD database. The predictive ability of the RWR algorithm was proved to be excellent based on the leave-one-out

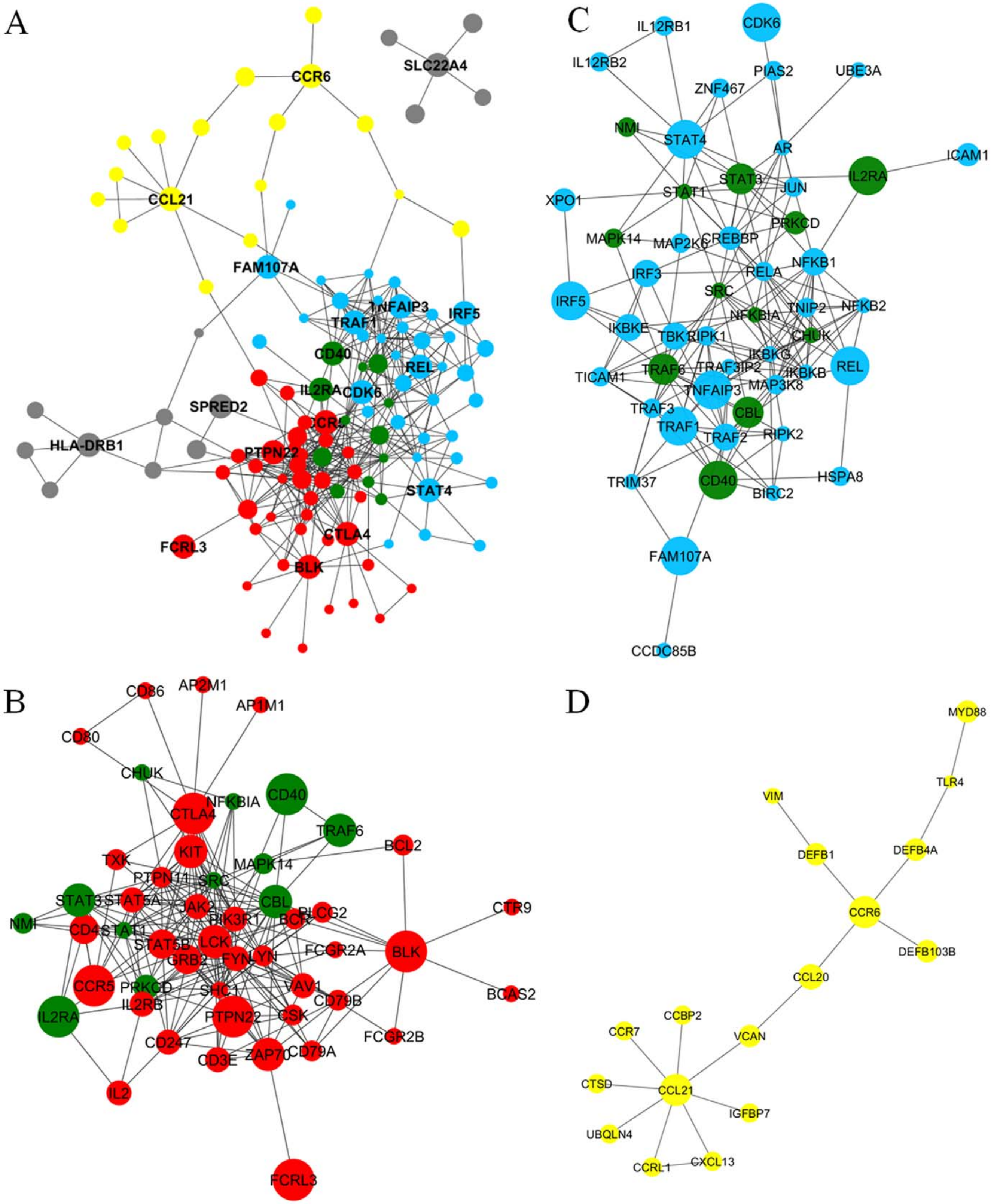

Figure 6. RA-associated network. A) Entire RA-associated network comprising known RA-associated genes and genes ranked in the top 100 by the RWR algorithm and edges are physical interactions between their products. Nodes are color coded by hierarchical clusters detected by the EAGLE algorithm: CL1, red; CL2; cyan, and CL3, yellow. Overlapped region between CL1 and CL2 are rendered in green. Node size is based on the ranking in the RWR algorithm. Official gene symbols are shown for known RA-associated genes. B–D) Subnetworks corresponds to the hierarchical clusters CL1-3.
doi:10.1371/journal.pone.0025389.g006

Table 3. Top-ranked GO and KEGG annotations for three clusters in RA-associated network.

Annotation	Term[A]	Count[B]	%[C]	FE[D]	P-value
Cluster 1					
GO:0045321	Leukocyte activation	20	40.0	23.3	1.4×10^{-21}
GO:0002521	Leukocyte differentiation	15	30.0	32.3	8.0×10^{-18}
hsa04660	T cell receptor signaling pathway	15	30.0	15.7	1.1×10^{-13}
GO:0006468	Protein amino acid phosphorylation	19	38.0	8.0	2.8×10^{-12}
Cluster 2					
hsa04620	Toll-like receptor signaling pathway	20	40.0	18.2	2.0×10^{-20}
hsa04622	RIG-I-like receptor signaling pathway	14	28.0	21.7	7.5×10^{-15}
GO:0007249	I-kappaB kinase/NF-kappaB cascade	12	24.0	31.6	4.3×10^{-14}
hsa05200	Pathways in cancer	20	40.0	5.4	2.0×10^{-10}
hsa04623	Cytosolic DNA-sensing pathway	10	20.0	22.0	2.0×10^{-10}
hsa04621	NOD-like receptor signaling pathway	11	22.0	15.1	8.7×10^{-10}
Cluster 3					
GO:0006935	Chemotaxis	9	52.9	47.6	1.9×10^{-12}
GO:0007626	Locomotory behavior	9	52.9	27.8	1.5×10^{-10}
GO:0006955	Immune response	11	64.7	13.5	2.7×10^{-10}
GO:0006952	Defense response	10	58.8	13.7	3.1×10^{-9}
GO:0019957	C-C chemokine binding	4	23.5	231.8	4.4×10^{-7}
GO:0016493	C-C chemokine receptor activity	4	23.5	231.8	4.4×10^{-7}

[A] Within each cluster, related terms are not shown to reduce redundancy. Among terms with parent-child relationships, we selected one showing highest significance enrichment P-value.

[B] Number of GO or KEGG category genes in each cluster.

[C] Percentage of GO or KEGG category genes in each cluster.

[D] Fold Enrichment of genes in each cluster compared to a background list.

doi:10.1371/journal.pone.0025389.t003

cross-validation by omitting each RA-associated gene (AUC = 84.4%). This result suggests that the RA-associated genes are in proximity to each other within the PPI network, which is consistent with the recent study showing that the products of RA-associated genes are more interconnected than would be expected by chance [94]. The top-ranked genes with the RWR algorithm are intriguing. The gene in the first position (*ZAP70*) is a causal gene of RA-like autoimmune arthritis in mice [78]. Furthermore, recessive and compound heterozygous mutations in *ZAP70* cause human severe combined immunodeficiency [95–97]. Within the top 100 ranked genes, genes implicating susceptibility to RA (*CD247*, *IL2RB* and *IL2*) and to other autoimmune diseases (*CD80*, *FCGR2A*, *FCGR2B*, *ICAM1*, *JAK2*, *LYN*, *NFKBIA*, *PTPN11*, *STAT3* and *TRAF3IP2*) are enriched.

We also found that the analysis of network topology was useful to find functional modules involved in the disease pathology. This method has two steps: first, a disease-related network comprising genes in the vicinity to curated susceptibility genes within the PPI network is constructed using the RWR algorithm. Second, the overlapping and hierarchical structure of the disease-related network is explored and functional annotation is implemented for each cluster. The systems genetics approach proposed here will be applicable to most common diseases and will work well especially when genes associated with the disease of interest are in proximity to each other within the PPI network. When applying the method to RA, the resulting clusters were fitted into different categories of immunological pathways.

CL1 is related to "leukocyte activation and differentiation" (Figure 6B, Table 3). T-cell differentiation plays an important role in autoimmunity. Strongly self-reactive T cells are primarily eliminated in the thymus by negative selection (central tolerance). Some of the self-reactive T cells, however, may escape from negative selection and can cause autoimmune diseases. Defect in thymic T-cell selection due to a mutation of *Zap70* causes autoimmune arthritis in mice [78]. CL2 fits into "pattern-recognition receptor signaling pathways" (Figure 6C, Table 3). The innate immune functions of macrophages and neutrophils depend on pattern-recognition receptors such as Toll-like receptors and Nod-like receptors. Genes relevant to these pattern-recognition receptor signaling pathways were enriched in CL2. CL3 corresponds to "chemokines and their receptors" (Figure 6D, Table 3). The main function shared by chemokines and chemokine receptors is leukocyte chemotaxis, which helps direct migration of leukocytes to an injury site. Genetic defects in these biological pathways can inappropriately activate immune cells leading to inflammation and host cell destruction that can cause autoimmune diseases. Notably, the clusters inferred from this study are similar to the pathways implicated by Zhernakova et al. [83], in which genes associated with autoimmune diseases are grouped into four categories: 'T cell differentiation', 'immune-cell activation and signaling', 'innate immunity and TNF signaling', and 'cytokines and chemokines'.

Some limitations of our study should be noted. Our electronic database search had been performed a year ago (on June 18 2010). Continuing efforts to examine whether more updated genetic findings appear in publication to renew the list of susceptibility genes to RA is required. By searching for the NHGRI database deposited after June 18 2010, we found four relevant articles.[80,98–100] Genetic variants outside the MHC region passing the genome-wide significant threshold (5.0×10^{-8}) were retrieved.

Two Asian GWASs detected SNPs on two genetic loci (*AIRE*, and *PADI4*) [98,99]. European GWAS identified 8 loci as shared genetic factors between RA and celiac disease [80]. We reconsidered our network analysis by including newly discovered loci. We could assign these 10 loci to 8 unique genes: *AIRE*, *PADI4*, *TRAF1*, *STAT4*, *YDJC*, *UBASH3A*, *CD247*, and *ATXN2*. *TRAF1*, *STAT4*, and *PADI4* were already included into our model. *YDJC* is not deposited in the HPRD database. Thus, 4 genes (*AIRE*, *UBASH3A*, *CD247*, and *ATXN2*) were newly included. We re-examined the RWR algorithm using a total of 23 genes as initial vertices. The rank correlation in genes between before and after including the 4 genes was 0.978. This indicates the ranking of candidate genes were not largely affected. The result from the hierarchical clustering method and functional annotation rendered similar biological pathways (Figure S4, Table S11). As previously mentioned, *CD247* was one of top-ranked genes (32th) in the original RWR analysis. The result of network analysis by using additional 4 newly discovered genes converged on the same biological pathways, suggesting the strong relevance of these pathways to the etiology of RA. Only physical PPI data were used to construct the molecular network and it is inevitably noisy and incomplete. A PPI network integrated with a transcription profiling network could improve the predictive ability of network-guided prioritization of genes [22,101].

We have demonstrated that recent successful discoveries of genetic variants associated with diseases are valuable resources to provide targets for future resequencing studies to reveal the biological pathways. Such efforts utilizing GWAS discoveries will accelerate genetic discoveries and improve the predictive ability of the genetic variations. While exploration of other types of genomic variation such as rare and low frequency single nucleotide changes and insertions and deletions of nucleotides is promising, it may be challenging because the variants are likely to be population-specific. The biological pathways highlighted by the various common genetic variants associated with the disease across populations will encourage examination and functional annotation of newly discovered rare variants.

Materials and Methods

Ethics statement

The Ethics Committee of Tokai University approved the study protocols and all participants gave written informed consent.

Study participants

1,287 RA subjects and 1,500 control subjects of Japanese origin were recruited. All cases were diagnosed by board certified rheumatologists and fulfilled 1987 American College of Rheumatology criteria [102]. The dataset was updated from our previous study [103].

Information on the positivity of anti-CCP and RF for 481 and 462 cases, respectively, was measured. Anti-CCP antibody titers were measured with the second generation ELISA kit (MESACUP CCP; Medical & Biological Laboratories Co. Ltd, Nagoya, Japan). A cut-off value of 4.5 U/ml was used for anti-CCP antibody positivity. RF positivity was determined by using N-Assay TIA RF Nittobo (Nitto Boseki Co., Ltd, Koriyama, Japan). The positivity of anti-CCP and RF was observed in 90.4% and 80.5% cases, respectively.

Genotyping

All study participants were genotyped for *HLA-DRB1* alleles and selected SNPs described below. Genotyping of *HLA-DRB1* alleles was performed by Luminex Multi-Analyte Profiling system (xMAP) with a WAKFlow HLA typing kit (Wakunaga, Hiroshima, Japan). Genotyping of SNPs was performed by TaqMan SNP Genotyping Assays on the ABI PRISM 7900HT Sequence Detection System (Applied Biosystems, Tokyo, Japan). Departure from Hardy-Weinberg equilibrium (HWE) in control samples was examined at the significance level of $P<0.001$ by means of the exact test using PLINK software [104].

Electronic database search strategies

PubMed search. We identified published meta-analyses addressing the association between genetic variants and RA risk in population-based studies. We performed a literature search of the PubMed database (last search June 18, 2010). Searches were conducted using the following keywords: "rheumatoid arthritis" and [genetic(s) or polymorphism(s) or allele(s) or mutation(s)] and (meta-analysis or metaanalysis or "systematic review"). Reference mining of retrieved articles was used to identify additional articles. Meta-analyses included in our analysis had to meet all of the following criteria: evaluated RA risk as the outcome (analyses of pharmacogenomics and RA severity were excluded) and published in English. Two researchers (HN and TC) conducted literature searches independently, and any disagreement between the two researchers was accommodated by the third researcher (AT).

Genetic models and methods for combining studies examined in the retrieved meta-analyses differed according to article. In order to evaluate evidence of association of genetic variant with RA in the same statistical manner, we performed re-analysis of published meta-analyses. Therefore, we included the following meta-analyses: greater than or equal to five independent studies were included and the total number of cases and controls was larger than 3,000; adequate data to calculate OR for each of the included studies was provided; and per-allele effect of risk allele was examined. When there were several meta-analyses on the same variant including different studies, we created a comprehensive set of individual studies using following criteria: i) studies did not overlap data between studies, and ii) study with largest sample size was used when studies overlapped data.

The NHGRI GWAS catalog. Recent update of GWAS findings was sought by the NHGRI GWAS catalog (http://www.genome.gov/gwastudies/ Accessed June 18, 2010). We included the associations that met the genome-wide significance of $P<5.0\times10^{-8}$ in our analysis.

Re-analysis of published meta-analyses and selection of genetic variants

In the re-analysis of all the retrieved meta-analyses, the per-allele ORs for individual studies were combined using both fixed effects model and Dersimonian-Laird random effects model meta-analyses. We examined the test for association at the significance level of $P<2.5\times10^{-3}$ ($=0.05/20$) to correct the multiple testing. Homogeneity across studies was examined by Cochran's Q test at the significance level of 0.1. The extent of between-study heterogeneity was quantified by I^2. I^2 values over 50% indicate large heterogeneity. All the meta-analyses were performed by using STATA version 11.0.

We selected genetic variants according to the following criteria: First, genetic variants showed evidence of association in the re-analysis of meta-analysis ($P<2.5\times10^{-3}$) or in the NHGRI GWAS catalog ($P<0.5\times10^{-8}$). Second, minor allele frequency was larger than 5% in Japanese as Janssens et al. shows low-frequency genetic variants with small effects does not largely affect the predictive ability [105]. The allele frequency in Japanese population was sought in SNP Control Database [106].

In order to introduce possible genetic heterogeneity among ethnic groups into the risk prediction model, we performed subgroup analyses in which ORs per ethnic group were estimated. We used the ethnic group-specific effect into risk prediction model if meta-analysis fulfilled the following criteria: greater than or equal to three independent studies were included and the total number of cases and controls was larger than 2,000 in both target (East Asian) and major (European descent) ethnic groups; the ethnic group-specific OR was statistically significant ($P<2.5\times10^{-3}$); and heterogeneity in the ethnic group-specific OR between the target and major ethnic groups was statistically significant ($P<0.05$).

Genetic risk models

We considered three logistic regression models: the HLA model included the *HLA-DRB1* alleles only; the non-HLA model included the selected genetic variants at the non-HLA loci; and the integrative model incorporated both of the *HLA-DRB1* alleles and the genetic variants at the non-HLA loci. In the logistic regression analyses, the genetic risk score is as follows:

$$GR(\mathbf{X},\mathbf{Z})=\sum\nolimits_{i=1}^{L} X_i \log OR_i+\sum\nolimits_{j=1}^{M} Z_j \log OR_j, \qquad (1)$$

where $X_i \in \{0,1,2\}$ is the number of risk alleles of SNP locus i, $\mathbf{X}=(X_1,\ldots,X_L)$ is the genetic profiles of L loci genotypes, $Z_j \in \{0,1,2\}$ is the indicator variable, indexing the number of each of selected *HLA-DRB1* alleles, $\mathbf{Z}=(Z_1,\ldots,Z_M)$ is the profiles showing the subject's *HLA-DRB1* genotype, and the OR for each variant is derived from the re-analysis of meta-analyses. The integrative model was the full model that was expressed as the equation [1], whereas the HLA and the non-HLA models were the reduced models where the first and second terms on the right-hand side of the equation [1] were excluded, respectively.

In order to assess the predictive ability of the models, we used the ROC curve and calculated the AUC [16]. By definition, the AUC is the probability that a randomly selected subject with the disease of interest has a higher score than a randomly selected subject without the disease. When the ROC analyses were implemented, we restricted analyses to subjects with complete genotype data. Thus, 1,231 cases and 1,445 controls were available. We compared the fits of the three models with Akaike's information criterion.

Simulation study

In most common diseases, the predictive ability of common genetic variants may be suboptimal at the current moment. We therefore performed a simulation study to address how many additional loci should be mapped to establish an acceptable level of genetic risk prediction (AUC = 80.0%).

We assumed two scenarios of allelic architecture of as-yet-discovered genetic variants. First, we assumed the common disease-common variant hypothesis, in which a large proportion of the missing heritability can be explained by common variants [107]. In this model, the per-allele OR was set to be 1.1 or 1.2 and the RAF was set to be 0.1 or 0.3. Second, we assumed that the multiple rare variants with intermediate effects remain undiscovered [14]. In this model, we assumed that the per-allele OR was 2.0 or 3.0 with RAF of 0.01.

To simulate the distribution of RA risks in the general population, we considered the constrained multiplicative model [108]. First, we set 'base model', where all the possible combinations of genotypes of *HLA-DRB1* alleles and selected SNPs were included. For the base model, the ORs derived from our case-control association study were used and the allele frequencies in Japanese were obtained from SNP Control Database [106] (shown in Text S1B). Next, we added N diallelic loci to the base model. For simplicity, we assumed that the frequency and the effect size of the risk allele at each additional locus are the same as p and OR, respectively. We assumed that each locus is both in HWE and in linkage equilibrium. We denote K as the prevalence of RA and set it to 0.01. Under the rare disease assumption, the relative risk can be approximated by the odds ratio. The risk and the joint probability of multi-locus genotype can be written as the product across loci:

$$\mathrm{g}(\mathbf{X},\mathbf{Z},\mathbf{W})=\mathrm{b}\,\Pi_{\mathrm{i}=1}^{\mathrm{L}}\,\mathrm{OR}_{\mathrm{i}}^{\mathrm{X_i}}\,\Pi_{\mathrm{j}=1}^{\mathrm{M}}\,\mathrm{OR}_{\mathrm{j}}^{\mathrm{Z_j}}\,\Pi_{\mathrm{k}=1}^{\mathrm{MN}}\,\mathrm{OR}_{\mathrm{k}}^{\mathrm{W_k}}, \text{ and} \qquad (2)$$

$$\mathrm{p}(\mathbf{X},\mathbf{Z},\mathbf{W})=\Pi_{\mathrm{i}=1}^{\mathrm{L}}\,\mathrm{p}(\mathrm{X_i})\,\Pi_{\mathrm{j}=1}^{\mathrm{M}}\,\mathrm{p}(\mathrm{Z_j})\,\Pi_{\mathrm{k}=1}^{\mathrm{N}}\,\mathrm{p}(\mathrm{W_k}),$$

respectively, where b is the background risk so that $E(g)=K$, $p(X_i)$ is the probabilities of X_i, and $\mathbf{W}=(W_1,\ldots,W_N)$ is the genetic profiles of N-locus genotypes with $W_k \in \{0,1,2\}$ representing the number of risk alleles of additional locus k. By the assumption that each additional locus has the same p and OR, the equation [2] can be written down as:

$$g(\mathbf{X},\mathbf{Z},\mathbf{W})=b\,\Pi_{i=1}^{L}\,OR_i^{X_i}\,\Pi_{j=1}^{M}\,OR_j^{Z_j}\,OR^{s}, \text{ and} \qquad (3)$$

$$p(\mathbf{X},\mathbf{Z},\mathbf{W})=\Pi_{i=1}^{L}\,p(X_i)\,\Pi_{j=1}^{M}\,p(Z_j)\binom{2N}{s}p^{s}(1-p)^{2N-s},$$

where $s=\sum_{k=1}^{N} W_k$. For some genetic profiles with many risk alleles, the risk expressed as the equation [3] can exceed 1. In the constrained multiplicative model, if the risk exceeds 1, the risk is set to 1 [108].

The probability of multi-locus genotype given disease status is:

$$p(\mathbf{X},\mathbf{Z},\mathbf{W}|Disease)=g(\mathbf{X},\mathbf{Z},\mathbf{W})\times p(\mathbf{X},\mathbf{Z},\mathbf{W})/\sum\nolimits_{\mathbf{X},\mathbf{Z},\mathbf{W}} g(\mathbf{X},\mathbf{Z},\mathbf{W})\times p(\mathbf{X},\mathbf{Z},\mathbf{W}), \text{ and}$$

$$p(\mathbf{X},\mathbf{Z},\mathbf{W}|NonDisease)=[p(\mathbf{X},\mathbf{Z},\mathbf{W})-K\times p(\mathbf{X},\mathbf{Z},\mathbf{W}|Disease)]/[1-K]. \qquad (4)$$

For an arbitrary cut-off value of t, the true and false positive rates are:

$$TPR(t)=\sum\nolimits_{\mathbf{X},\mathbf{Z},\mathbf{W}:g(\mathbf{X},\mathbf{Z},\mathbf{W})\geq t} p(\mathbf{X},\mathbf{Z},\mathbf{W}|Disease), \text{ and}$$

$$FPR(t)=\sum\nolimits_{\mathbf{X},\mathbf{Z},\mathbf{W}:g(\mathbf{X},\mathbf{Z},\mathbf{W})\geq t} p(\mathbf{X},\mathbf{Z},\mathbf{W}|NonDisease),$$

respectively.

Given the TPR and FPR at each cut-off value t, the ROC curve can be drawn and then the AUC can be calculated by the trapezoid rule [109].

Network analysis

We assigned selected genetic variants to a single protein-coding gene according to the following hierarchy: coding>intronic>5′UTR>3′UTR>near gene (within 2 kb to 5′ or 0.5 kb to 3′ of a gene)>intergenic. If a selected variant mapped an intergenic region, we sought literature of fine-mapping studies or GWAS of RA and other autoimmune diseases showing evidence of association of variants in higher levels of the hierarchy.

The physical PPI network was constructed using the HPRD database [75,76]. In PPI networks, vertices are proteins and edges represent a physical interaction between two proteins. We projected the RA-associated genes onto the constructed PPI network and candidate genes were then ranked based on the global distance to the RA-associated genes within the PPI network by using random walk with restart (RWR) algorithm [77]. The RWR algorithm is a powerful tool to measure proximity between vertices on complex network.

In a random walk, starting at some initial 'seed' vertices (i.e., proteins encoded by the RA-associated genes), we chose at random an edge that is attached to the current vertex and move along the chosen edge to the linked vertex, and iterate many steps. In the RWR, at each step of the walk we return to the initial seed vertices with the restart probability, r. All vertices are ranked by the number of times that the walker visits to corresponding vertices in the process. The outline is described below.

The adjacency matrix $\mathbf{A}$ of the PPI network is the matrix with elements A_{ij} as follows: $A_{ij}=\begin{cases}1 & \text{if there is an edge between vertices } i \text{ and } j\\ 0 & \text{otherwise}\end{cases}$. We define the transition probability matrix $\mathbf{M}$ so that the transition probability M_{ij} from protein i to protein j is: $M_{ij}=A_{ij}\Big/\sum_j A_{ij}$. Let $\mathbf{p}^{(t)}$ be a vector whose i-th element holds the probability of a random walker being at vertex i at step t and $\mathbf{p}^{(0)}$ be the initial-state probability vector, the probability vector at the step t+1 is as follows:

$$\mathbf{p}^{(t+1)}=(1-r)\mathbf{M}\mathbf{p}^{(t)}+r\mathbf{p}^{(0)}.$$

In this study, $\mathbf{p}^{(0)}$ was defined as the vector with elements: $p_i^{(0)}=\begin{cases}1/\text{number of RA-associated genes} & \text{if vertex } i \text{ is RA-associated gene}\\ 0 & \text{otherwise}\end{cases}$. The restart probability r was set to be 0.5. We considered the random walker reached a steady-state when the difference between $\mathbf{p}^{(t+1)}$ and $\mathbf{p}^{(t)}$ (measured by the L1 norm) reached 10^{-10}. All the genes in the PPI network were ranked according to the corresponding values in the steady-state probability vector $\mathbf{p}^{(\infty)}$.

The predictive ability of the network-guided prioritization of genes was tested using leave-one-out cross-validation by omitting each RA-associated gene in turn from initial 'seed' vertices and performing the RWR algorithm for the purpose of its own evaluation. The ROC curve was drawn by plotting the TPR versus the FPR for all genes ranked above a sliding ranking threshold.

We define RA-associated network as a subnetwork in which vertices are the RA-associated genes and genes ranked in the top 100 by the RWR algorithm and edges are physical interactions between their products. Functional modules are then explored in the RA-associated network. The overlapping and hierarchical clusters were detected by using the EAGLE algorithm [84]. The functional annotation for the retrieved clusters was performed by using DAVID [85,86]. We set 9,521 genes on the PPI network from HPRD as the background in enrichment analysis.

Supporting Information

Figure S1 Flowchart detailing the exclusion and inclusion criteria and the number of studies excluded and included at each step of the electronic database searches. A) PubMed, and B) NHGRI GWAS catalog.
(TIF)

Figure S2 RA-associated network comprising known RA-associated genes and genes ranked in the top 50 by the RWR algorithm and edges are physical interactions between their products. Nodes are color coded by hierarchical clusters detected by the EAGLE algorithm: CL1, red; CL2; cyan, and CL3, yellow. Overlapped regions between CL1 and CL2 are rendered in green. Node size is based on the ranking in the RWR algorithm.
(TIF)

Figure S3 RA-associated network comprising known RA-associated genes and genes ranked in the top 150 by the RWR algorithm and edges are physical interactions between their products. Nodes are color coded by hierarchical clusters detected by the EAGLE algorithm: CL1, red; CL2; cyan, CL3, yellow; and CL4, orange. Overlapped regions between CL1 and CL2, CL1 and CL4, and CL2 and CL4 are rendered in green, pink, and purple, respectively. Node size is based on the ranking in the RWR algorithm.
(TIF)

Figure S4 Re-consideration on RA-associated network. The RWR algorithm was re-examined by adding recently discovered 4 genes (*AIRE*, *CD247*, *UBASH3A*, and *ATXN2*). Nodes are color coded by hierarchical clusters detected by the EAGLE algorithm: CL1, red; CL2; cyan, and CL3, yellow. Overlapped regions between CL1 and CL2 are rendered in green. Node size is based on the ranking in the RWR algorithm.
(TIFF)

Table S1 Result of ratings for 87 abstracts retrieved from PubMed. The scoring was conducted by independent two authors (Hirofumi Nakaoka and Tailin Cui), which is color-coded in green and blue, respectively. Any disagreement between the two researchers was accommodated by Atsushi Tajima. The final decision is rendered in red.
(DOC)

Table S2 Result of rating for 54 full-text articles.
(DOC)

Table S3 Result of screening of extracted data from 51 full-text articles.
(DOC)

Table S4 Re-analysis of meta-analyses addressing genetic associations with RA risk.
(DOC)

Table S5 Assignment of a single gene to genetic variants associated with RA and the allele frequencies in European and Japanese.
(DOC)

Table S6 Ethnic group-specific analysis of published meta-analyses of genetic associations with RA risk. The SNPs in which the heterogeneity in the ORs between European and East Asian populations are significant are highlighted in yellow.
(DOC)

Table S7 Genotype counts for six *HLA-DRB1* alleles and 15 SNPs.
(DOC)

Table S8 Association analysis of RF and anti-CCP positive RA patients versus control subjects with selected genetic variants.
(DOC)

Table S9 GO and KEGG annotations for three clusters in RA-associated network comprising RA-associated genes and genes ranked in the top 50 by the RWR algorithm.
(DOC)

Table S10 GO and KEGG annotations for three clusters in RA-associated network comprising RA-associated genes and genes ranked in the top 150 by the RWR algorithm.
(DOC)

Table S11 Re-consideration on RA-associated network. The RWR algorithm was re-examined by adding recently discovered 4 genes (*AIRE*, *CD247*, *UBASH3A*, and *ATXN2*). GO and KEGG annotations for three clusters in RA-associated network comprising RA-associated genes and genes ranked in the top 100 by the RWR algorithm.
(DOC)

Text S1 Supplementary methods.
(DOC)

Acknowledgments

We thank all the study participants and supporting medical staffs for making this study possible. We are grateful to Keiko Asami and Hiromi Moriya for their contribution to sample data collection. We wish to thank Hiromi Kamura, Kaori Fukushima, Ryota Sugimoto and Hideki Hayashi for their technical support.

Author Contributions

Conceived and designed the experiments: HN AT HI II. Performed the experiments: AO SM KK. Analyzed the data: HN TC AT. Contributed reagents/materials/analysis tools: YH SS YS HI. Wrote the paper: HN AT II.

References

1. Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118(5): 1590–1605.
2. Hirschhorn JN (2009) Genomewide association studies–illuminating biologic pathways. N Engl J Med 360(17): 1699–1701. 10.1056/NEJMp0808934.
3. Yang Q, Khoury MJ, Botto L, Friedman JM, Flanders WD (2003) Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. Am J Hum Genet 72(3): 636–649. 10.1086/367923.
4. Gulcher J, Stefansson K (2010) Genetic risk information for common diseases may indeed be already useful for prevention and early detection. Eur J Clin Invest 40(1): 56–63. 10.1111/j.1365-2362.2009.02233.x.
5. Janssens AC, van Duijn CM (2008) Genome-based prediction of common diseases: Advances and prospects. Hum Mol Genet 17(R2): R166–73. 10.1093/hmg/ddn250.
6. Kraft P, Hunter DJ (2009) Genetic risk prediction–are we there yet? N Engl J Med 360(17): 1701–1703. 10.1056/NEJMp0810107.
7. Pharoah PD, Antoniou AC, Easton DF, Ponder BA (2008) Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med 358(26): 2796–2803. 10.1056/NEJMsa0708739.
8. Ransohoff DF, Khoury MJ (2010) Personal genomics: Information can be harmful. Eur J Clin Invest 40(1): 64–68. 10.1111/j.1365-2362.2009.02232.x.
9. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461(7265): 747–753. 10.1038/nature08494.
10. Goldstein DB (2009) Common genetic variation and human traits. N Engl J Med 360(17): 1696–1698. 10.1056/NEJMp0806284.
11. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, et al. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305(5685): 869–872. 10.1126/science.1099870.
12. Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, et al. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40(5): 592–599. 10.1038/ng.118.
13. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324(5925): 387–389. 10.1126/science.1167728.
14. McCarthy MI (2009) Exploring the unknown: Assumptions about allelic architecture and strategies for susceptibility variant discovery. Genome Med 1(7): 66. 10.1186/gm66.
15. Janssens AC, Gwinn M, Bradley LA, Oostra BA, van Duijn CM, et al. (2008) A critical appraisal of the scientific basis of commercial genomic profiles used to assess health risks and personalize health interventions. Am J Hum Genet 82(3): 593–599. 10.1016/j.ajhg.2007.12.020.
16. Cook NR, Ridker PM (2009) Advances in measuring the effect of individual predictors of cardiovascular risk: The role of reclassification measures. Ann Intern Med 150(11): 795–802.
17. Janssens AC, van Duijn CM (2009) Genome-based prediction of common diseases: Methodological considerations for future research. Genome Med 1(2): 20. 10.1186/gm20.
18. NCI-NHGRI Working Group on Replication in Association Studies, Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, et al. (2007) Replicating genotype-phenotype associations. Nature 447(7145): 655–660.
19. Khoury MJ, Gwinn M, Ioannidis JP (2010) The emergence of translational epidemiology: From scientific discovery to population health impact. Am J Epidemiol 172(5): 517–524. 10.1093/aje/kwq211.
20. Nakaoka H, Inoue I (2009) Meta-analysis of genetic association studies: Methodologies, between-study heterogeneity and winner's curse. J Hum Genet 54(11): 615–623. 10.1038/jhg.2009.95.
21. Nadeau JH, Dudley AM (2011) Genetics. systems genetics. Science 331(6020): 1015–1016. 10.1126/science.1203869.
22. Vidal M, Cusick ME, Barabasi AL (2011) Interactome networks and human disease. Cell 144(6): 986–998. 10.1016/j.cell.2011.02.016.
23. Klareskog L, Catrina AI, Paget S (2009) Rheumatoid arthritis. Lancet 373(9664): 659–672. 10.1016/S0140-6736(09)60008-8.
24. Orozco G, Barton A (2010) Update on the genetic risk factors for rheumatoid arthritis. Expert Rev Clin Immunol 6(1): 61–75.
25. Plenge RM (2009) Recent progress in rheumatoid arthritis genetics: One step towards improved patient care. Curr Opin Rheumatol 21(3): 262–271. 10.1097/BOR.0b013e32832a2e2d.
26. van der Helm-van Mil AH, Huizinga TW (2008) Advances in the genetics of rheumatoid arthritis point to subclassification into distinct disease subsets. Arthritis Res Ther 10(2): 205. 10.1186/ar2384.
27. Begovich AB, Chang M, Schrodi SJ (2007) Meta-analysis evidence of a differential risk of the FCRL3 -169T→C polymorphism in white and east asian rheumatoid arthritis patients. Arthritis Rheum 56(9): 3168–3171. 10.1002/art.22857.
28. Burr ML, Naseem H, Hinks A, Eyre S, Gibbons LJ, et al. (2010) PADI4 genotype is not associated with rheumatoid arthritis in a large UK caucasian population. Ann Rheum Dis 69(4): 666–670. 10.1136/ard.2009.111294.
29. Daha NA, Kurreeman FA, Marques RB, Stoeken-Rijsbergen G, Verduijn W, et al. (2009) Confirmation of STAT4, IL2/IL21, and CTLA4 polymorphisms in rheumatoid arthritis. Arthritis Rheum 60(5): 1255–1260. 10.1002/art.24503.
30. Dieguez-Gonzalez R, Calaza M, Perez-Pampin E, de la Serna AR, Fernandez-Gutierrez B, et al. (2008) Association of interferon regulatory factor 5 haplotypes, similar to that found in systemic lupus erythematosus, in a large subgroup of patients with rheumatoid arthritis. Arthritis Rheum 58(5): 1264–1274. 10.1002/art.23426.
31. Han S, Li Y, Mao Y, Xie Y (2005) Meta-analysis of the association of CTLA-4 exon-1 +49A/G polymorphism with rheumatoid arthritis. Hum Genet 118(1): 123–132. 10.1007/s00439-005-0033-9.
32. Han SW, Lee WK, Kwon KT, Lee BK, Nam EJ, et al. (2009) Association of polymorphisms in interferon regulatory factor 5 gene with rheumatoid arthritis: A metaanalysis. J Rheumatol 36(4): 693–697. 10.3899/jrheum.081054.
33. Harrison P, Pointon JJ, Chapman K, Roddam A, Wordsworth BP (2008) Interleukin-1 promoter region polymorphism role in rheumatoid arthritis: A meta-analysis of IL-1B-511A/G variant reveals association with rheumatoid arthritis. Rheumatology (Oxford) 47(12): 1768–1770. 10.1093/rheumatology/ken374.
34. Iwamoto T, Ikari K, Nakamura T, Kuwahara M, Toyama Y, et al. (2006) Association between PADI4 and rheumatoid arthritis: A meta-analysis. Rheumatology (Oxford) 45(7): 804–807. 10.1093/rheumatology/kel023.
35. Ji JD, Lee WJ, Kong KA, Woo JH, Choi SJ, et al. (2010) Association of STAT4 polymorphism with rheumatoid arthritis and systemic lupus erythematosus: A meta-analysis. Mol Biol Rep 37(1): 141–147. 10.1007/s11033-009-9553-z.
36. Lee YH, Ji JD, Song GG (2009) Association between interleukin 1 polymorphisms and rheumatoid arthritis susceptibility: A metaanalysis. J Rheumatol 36(1): 12–15. 10.3899/jrheum.080450.

37. Lee YH, Ji JD, Song GG (2008) Associations between FCGR3A polymorphisms and susceptibility to rheumatoid arthritis: A metaanalysis. J Rheumatol 35(11): 2129–2135.
38. Lee YH, Ji JD, Song GG (2007) Tumor necrosis factor-alpha promoter -308 A/G polymorphism and rheumatoid arthritis susceptibility: A metaanalysis. J Rheumatol 34(1): 43–49.
39. Lee YH, Rho YH, Choi SJ, Ji JD, Song GG (2007) PADI4 polymorphisms and rheumatoid arthritis susceptibility: A meta-analysis. Rheumatol Int 27(9): 827–833. 10.1007/s00296-007-0320-y.
40. Lee YH, Rho YH, Choi SJ, Ji JD, Song GG, et al. (2007) The PTPN22 C1858T functional polymorphism and autoimmune diseases–a meta-analysis. Rheumatology (Oxford) 46(1): 49–56. 10.1093/rheumatology/kel170.
41. Lee YH, Woo JH, Choi SJ, Ji JD, Song GG (2010) Fc receptor-like 3 -169 C/T polymorphism and RA susceptibility: A meta-analysis. Rheumatol Int 30(7): 947–953. 10.1007/s00296-009-1082-5.
42. Lee YH, Woo JH, Choi SJ, Ji JD, Song GG (2010) Association between the rs7574865 polymorphism of STAT4 and rheumatoid arthritis: A meta-analysis. Rheumatol Int 30(5): 661–666. 10.1007/s00296-009-1051-z.
43. Lei C, Dongqing Z, Yeqing S, Oaks MK, Lishan C, et al. (2005) Association of the CTLA-4 gene with rheumatoid arthritis in chinese han population. Eur J Hum Genet 13(7): 823–828. 10.1038/sj.ejhg.5201423.
44. Lindner E, Nordang GB, Melum E, Flato B, Selvaag AM, et al. (2007) Lack of association between the chemokine receptor 5 polymorphism CCR5delta32 in rheumatoid arthritis and juvenile idiopathic arthritis. BMC Med Genet 8: 33. 10.1186/1471-2350-8-33.
45. Okada Y, Mori M, Yamada R, Suzuki A, Kobayashi K, et al. (2008) SLC22A4 polymorphism and rheumatoid arthritis susceptibility: A replication study in a japanese population and a metaanalysis. J Rheumatol 35(9): 1723–1728.
46. Orozco G, Alizadeh BZ, Delgado-Vega AM, Gonzalez-Gay MA, Balsa A, et al. (2008) Association of STAT4 with rheumatoid arthritis: A replication study in three european populations. Arthritis Rheum 58(7): 1974–1980. 10.1002/art.23549.
47. Orozco G, Eyre S, Hinks A, Ke X, Wellcome Trust Case Control consortium YEAR Consortium, et al. (2010) Association of CD40 with rheumatoid arthritis confirmed in a large UK case-control study. Ann Rheum Dis 69(5): 813–816. 10.1136/ard.2009.109579.
48. Patsopoulos NA, Ioannidis JP (2010) Susceptibility variants for rheumatoid arthritis in the TRAF1-C5 and 6q23 loci: A meta-analysis. Ann Rheum Dis 69(3): 561–566. 10.1136/ard.2009.109447.
49. Plant D, Barton A, Thomson W, Ke X, Eyre S, et al. (2009) A re-evaluation of three putative functional single nucleotide polymorphisms in rheumatoid arthritis. Ann Rheum Dis 68(8): 1373–1375. 10.1136/ard.2008.103572.
50. Plenge RM, Padyukov L, Remmers EF, Purcell S, Lee AT, et al. (2005) Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from north america and sweden: Association of susceptibility with PTPN22, CTLA4, and PADI4. Am J Hum Genet 77(6): 1044–1060. 10.1086/498651.
51. Prahalad S (2006) Negative association between the chemokine receptor CCR5-Delta32 polymorphism and rheumatoid arthritis: A meta-analysis. Genes Immun 7(3): 264–268. 10.1038/sj.gene.6364298.
52. Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, et al. (2008) Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat Genet 40(10): 1216–1223. 10.1038/ng.233.
53. Suarez-Gestal M, Calaza M, Dieguez-Gonzalez R, Perez-Pampin E, Pablos JL, et al. (2009) Rheumatoid arthritis does not share most of the newly identified systemic lupus erythematosus genetic factors. Arthritis Rheum 60(9): 2558–2564. 10.1002/art.24748.
54. Takata Y, Inoue H, Sato A, Tsugawa K, Miyatake K, et al. (2008) Replication of reported genetic associations of PADI4, FCRL3, SLC22A4 and RUNX1 genes with rheumatoid arthritis: Results of an independent japanese population and evidence from meta-analysis of east asian studies. J Hum Genet 53(2): 163–173. 10.1007/s10038-007-0232-4.
55. Wheeler J, McHale M, Jackson V, Penny M (2007) Assessing theoretical risk and benefit suggested by genetic association studies of CCR5: Experience in a drug development programme for maraviroc. Antivir Ther 12(2): 233–245.
56. Barnetche T, Constantin A, Cantagrel A, Cambon-Thomsen A, Gourraud PA (2008) New classification of HLA-DRB1 alleles in rheumatoid arthritis susceptibility: A combined analysis of worldwide samples. Arthritis Res Ther 10(1): R26. 10.1186/ar2379.
57. Delgado-Vega AM, Anaya JM (2007) Meta-analysis of HLA-DRB1 polymorphism in latin american patients with rheumatoid arthritis. Autoimmun Rev 6(6): 402–408. 10.1016/j.autrev.2006.11.004.
58. Ioannidis JP, Tarassi K, Papadopoulos IA, Voulgari PV, Boki KA, et al. (2002) Shared epitopes and rheumatoid arthritis: Disease associations in greece and meta-analysis of mediterranean european populations. Semin Arthritis Rheum 31(6): 361–370.
59. Jun KR, Choi SE, Cha CH, Oh HB, Heo YS, et al. (2007) Meta-analysis of the association between HLA-DRB1 allele and rheumatoid arthritis susceptibility in asian populations. J Korean Med Sci 22(6): 973–980.
60. Williams RC, Jacobsson LT, Knowler WC, del Puente A, Kostyu D, et al. (1995) Meta-analysis reveals association between most common class II haplotype in full-heritage native americans and rheumatoid arthritis. Hum Immunol 42(1): 90–94.
61. van der Woude D, Lie BA, Lundstrom E, Balsa A, Feitsma AL, et al. (2010) Protection against anti-citrullinated protein antibody-positive rheumatoid arthritis is predominantly associated with HLA-DRB1*1301: A meta-analysis of HLA-DRB1 associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in four european populations. Arthritis Rheum 62(5): 1236–1245. 10.1002/art.27366.
62. Fernando MM, Stevens CR, Walsh EC, De Jager PL, Goyette P, et al. (2008) Defining the role of the MHC in autoimmunity: A review and pooled analysis. PLoS Genet 4(4): e1000024. 10.1371/journal.pgen.1000024.
63. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4(2): 45–61.
64. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33(2): 177–182.
65. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145): 661–678.
66. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, et al. (2007) TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study. N Engl J Med 357(12): 1199–1209. 10.1056/NEJMoa073491.
67. Plenge RM, Cotsapas C, Davies L, Price AL, de Bakker PI, et al. (2007) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat Genet 39(12): 1477–1482. 10.1038/ng.2007.27.
68. Julia A, Ballina J, Canete JD, Balsa A, Tornero-Molina J, et al. (2008) Genome-wide association study of rheumatoid arthritis in the spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility. Arthritis Rheum 58(8): 2275–2286. 10.1002/art.23623.
69. Gregersen PK, Amos CI, Lee AT, Lu Y, Remmers EF, et al. (2009) REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. Nat Genet 41(7): 820–823. 10.1038/ng.395.
70. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, et al. (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet 42(6): 508–514. 10.1038/ng.582.
71. Kochi Y, Okada Y, Suzuki A, Ikari K, Terao C, et al. (2010) A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. Nat Genet 42(6): 515–519. 10.1038/ng.583.
72. Shimane K, Kochi Y, Horita T, Ikari K, Amano H, et al. (2010) The association of a nonsynonymous single-nucleotide polymorphism in TNFAIP3 with systemic lupus erythematosus and rheumatoid arthritis in the japanese population. Arthritis Rheum 62(2): 574–579. 10.1002/art.27190.
73. Janssens AC, Ioannidis JP, van Duijn CM, Little J, Khoury MJ, et al. (2011) Strengthening the reporting of genetic RIsk prediction studies: The GRIPS statement. PLoS Med 8(3): e1000420. 10.1371/journal.pmed.1000420.
74. Nishimura K, Sugiyama D, Kogata Y, Tsuji G, Nakazawa T, et al. (2007) Meta-analysis: Diagnostic accuracy of anti-cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis. Ann Intern Med 146(11): 797–808.
75. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13(10): 2363–2371. 10.1101/gr.1680803.
76. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human protein reference database–2009 update. Nucleic Acids Res 37(Database issue): D767–72. 10.1093/nar/gkn892.
77. Kohler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82(4): 949–958. 10.1016/j.ajhg.2008.02.013.
78. Sakaguchi N, Takahashi T, Hata H, Nomura T, Tagami T, et al. (2003) Altered thymic T-cell selection due to a mutation of the ZAP-70 gene causes autoimmune arthritis in mice. Nature 426(6965): 454–460. 10.1038/nature02119.
79. Barton A, Thomson W, Ke X, Eyre S, Hinks A, et al. (2008) Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. Nat Genet 40(10): 1156–1159. 10.1038/ng.218.
80. Zhernakova A, Stahl EA, Trynka G, Raychaudhuri S, Festen EA, et al. (2011) Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet 7(2): e1002004. 10.1371/journal.pgen.1002004.
81. Zhernakova A, Alizadeh BZ, Bevova M, van Leeuwen MA, Coenen MJ, et al. (2007) Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. Am J Hum Genet 81(6): 1284–1288. 10.1086/522037.
82. Barton A, Eyre S, Ke X, Hinks A, Bowes J, et al. (2009) Identification of AF4/FMR2 family, member 3 (AFF3) as a novel rheumatoid arthritis susceptibility locus and confirmation of two further pan-autoimmune susceptibility genes. Hum Mol Genet 18(13): 2518–2522. 10.1093/hmg/ddp177.
83. Zhernakova A, van Diemen CC, Wijmenga C (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat Rev Genet 10(1): 43–55. 10.1038/nrg2489.
84. Shen H, Cheng X, Cai K, Hu MB (2009) Detect overlapping and hierarchical community structure in networks. Physica A 388(8): 1706–1712.

85. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1): 44–57. 10.1038/nprot.2008.211.

86. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1): 1–13. 10.1093/nar/gkn923.

87. Kochi Y, Suzuki A, Yamada R, Yamamoto K (2009) Genetics of rheumatoid arthritis: Underlying evidence of ethnic differences. J Autoimmun 32(3–4): 158–162. 10.1016/j.jaut.2009.02.020.

88. Kurreeman F, Liao K, Chibnik L, Hickey B, Stahl E, et al. (2011) Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. Am J Hum Genet 88(1): 57–69. 10.1016/j.ajhg.2010.12.007.

89. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11(6): 415–425. 10.1038/nrg2779.

90. Lee-Kirsch MA, Gong M, Chowdhury D, Senenko L, Engel K, et al. (2007) Mutations in the gene encoding the 3′-5′ DNA exonuclease TREX1 are associated with systemic lupus erythematosus. Nat Genet 39(9): 1065–1067. 10.1038/ng2091.

91. Surolia I, Pirnie SP, Chellappa V, Taylor KN, Cariappa A, et al. (2010) Functionally defective germline variants of sialic acid acetylesterase in autoimmunity. Nature 466(7303): 243–247. 10.1038/nature09115.

92. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, et al. (2008) Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods 5(10): 887–893. 10.1038/nmeth.1251.

93. Kenny EM, Cormican P, Gilks WP, Gates AS, O'Dushlaine CT, et al. (2011) Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. DNA Res 18(1): 31–38. 10.1093/dnares/dsq029.

94. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, et al. (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet 7(1): e1001273. 10.1371/journal.pgen.1001273.

95. Arpaia E, Shahar M, Dadi H, Cohen A, Roifman CM (1994) Defective T cell receptor signaling and CD8+ thymic selection in humans lacking zap-70 kinase. Cell 76(5): 947–958.

96. Chan AC, Kadlecek TA, Elder ME, Filipovich AH, Kuo WL, et al. (1994) ZAP-70 deficiency in an autosomal recessive form of severe combined immunodeficiency. Science 264(5165): 1599–1601.

97. Elder ME, Lin D, Clever J, Chan AC, Hope TJ, et al. (1994) Human severe combined immunodeficiency due to a defect in ZAP-70, a T cell tyrosine kinase. Science 264(5165): 1596–1599.

98. Freudenberg J, Lee HS, Han BG, Shin HD, Kang YM, et al. (2011) Genome-wide association study of rheumatoid arthritis in koreans: Population-specific loci as well as overlap with european susceptibility loci. Arthritis Rheum 63(4)): 884–893. 10.1002/art.30235; 10.1002/art.30235.

99. Terao C, Yamada R, Ohmura K, Takahashi M, Kawaguchi T, et al. (2011) The human AIRE gene at chromosome 21q22 is a genetic determinant for the predisposition to rheumatoid arthritis in japanese population. Hum Mol Genet 20(13): 2680–2685. 10.1093/hmg/ddr161.

100. Eleftherohorinou H, Hoggart CJ, Wright VJ, Levin M, Coin LJ (2011) Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. Hum Mol Genet 20(17): 3494–3506. 10.1093/hmg/ddr248.

101. Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, et al. (2005) Predictive models of molecular machines involved in caenorhabditis elegans early embryogenesis. Nature 436(7052): 861–865. 10.1038/nature03876.

102. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, et al. (1988) The american rheumatism association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 31(3): 315–324.

103. Tamiya G, Shinya M, Imanishi T, Ikuta T, Makino S, et al. (2005) Whole genome association study of rheumatoid arthritis using 27 039 microsatellites. Hum Mol Genet 14(16): 2305–2321. 10.1093/hmg/ddi234.

104. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3): 559–575.

105. Janssens AC, Moonesinghe R, Yang Q, Steyerberg EW, van Duijn CM, et al. (2007) The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. Genet Med 9(8): 528–535. 10.1097/GIM.0b013e31812eece0.

106. Koike A, Nishida N, Inoue I, Tsuji S, Tokunaga K (2009) Genome-wide association database developed in the japanese integrated database project. J Hum Genet 54(9): 543–546. 10.1038/jhg.2009.68.

107. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42(7): 565–569. 10.1038/ng.608.

108. Slatkin M (2008) Exchangeable models of complex inherited diseases. Genetics 179(4): 2253–2261. 10.1534/genetics.107.077719.

109. Lu Q, Elston RC (2008) Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. Am J Hum Genet 82(3): 641–651. 10.1016/j.ajhg.2007.12.025.