nature genetics

Common variants at *CD40* and other loci confer risk of rheumatoid arthritis

Soumya Raychaudhuri[1–3], Elaine F Remmers[4], Annette T Lee[5], Rachel Hackett[1], Candace Guiducci[1], Noël P Burtt[1], Lauren Gianniny[1], Benjamin D Korman[4], Leonid Padyukov[6], Fina A S Kurreeman[7], Monica Chang[8], Joseph J Catanese[8], Bo Ding[6], Sandra Wong[1], Annette H M van der Helm-van Mil[7], Benjamin M Neale[1,3,9], Jonathan Coblyn[2], Jing Cui[2], Paul P Tak[10], Gert Jan Wolbink[11,12], J Bart A Crusius[13], Irene E van der Horst-Bruinsma[14], Lindsey A Criswell[15], Christopher I Amos[16], Michael F Seldin[17], Daniel L Kastner[4], Kristin G Ardlie[1,18], Lars Alfredsson[19], Karen H Costenbader[2], David Altshuler[1,3], Tom W J Huizinga[7], Nancy A Shadick[2], Michael E Weinblatt[2], Niek de Vries[10], Jane Worthington[20], Mark Seielstad[21], Rene E M Toes[7], Elizabeth W Karlson[2], Ann B Begovich[8], Lars Klareskog[6],
Q1 Peter K Gregersen[5], Mark J Daly[1,3] & Robert M Plenge[1–3]

To identify rheumatoid arthritis risk loci in European populations, we conducted a meta-analysis of two published genome-wide association (GWA) studies totaling 3,393 cases and 12,462 controls[1,2]. We genotyped 31 top-ranked SNPs not previously associated with rheumatoid arthritis in an independent replication of 3,929 autoantibody-positive rheumatoid arthritis cases and 5,807 matched controls from eight separate collections. We identified a common variant at the *CD40* gene locus (rs4810485, $P = 0.0032$ replication, $P = 8.2 \times 10^{-9}$ overall, OR = 0.87). Along with other associations near *TRAF1* (refs. 2,3) and *TNFAIP3* (refs. 4,5), this implies a central role for the CD40 signaling pathway in rheumatoid arthritis pathogenesis. We also identified Q3 association at the *CCL21* gene locus (rs2812378, $P = 0.00097$ replication, $P = 2.8 \times 10^{-7}$ overall), a gene involved in lymphocyte trafficking. Finally, we identified evidence of association at four additional gene loci: *MMEL1-TNFRSF14* (rs3890745, $P = 0.0035$ replication, $P = 1.1 \times 10^{-7}$ overall), *CDK6* (rs42041, $P = 0.010$ replication, $P = 4.0 \times 10^{-6}$ overall), *PRKCQ* (rs4750316, $P = 0.0078$ replication, $P = 4.4 \times 10^{-6}$ overall), and *KIF5A-PIP4K2C* (rs1678542, $P = 0.0026$ replication, $P = 8.8 \times 10^{-8}$ overall). Q4

Rheumatoid arthritis is a systemic autoimmune disease with intra- Q5 articular inflammation as a dominant feature that affects up to 1% of the population. It can be subdivided clinically by the presence or absence of autoantibodies (antibodies to cyclic citrullinated peptide Q6 (CCP) or rheumatoid factor (RF), both of which are highly correlated to each other). Previous genetic studies have identified and validated five risk loci for autoantibody-positive RA: multiple alleles within the MHC region[6]; a missense allele in the *PTPN22* gene[7]; two alleles at the 6q23 locus near the *TNFAIP3* gene[4,5]; and single alleles in the *STAT4* locus[8] and *TRAF1-C5* loci[2]. Additional alleles at 4q27 (ref. 9), *CTLA4* (ref. 10) and *PADI4* (ref. 11) have suggestive associations, but have not yet been widely replicated in individuals of European ancestry.

To identify a collection of unbiased candidate rheumatoid arthritis risk loci for further investigation, we carried out a meta-analysis of SNP data from three case-control collections of European individuals from two published GWA studies[1,2] (**Table 1**, see Methods for details). We investigated a common set of ∼340,000 SNPs genotyped by the

[1]Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [2]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [3]Center for Human Genetic Research and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [4]Genetics and Genomics Branch, National Institute of Arthritis and Musculoskeletal and Skin Diseases, US National Institutes of Health, Bethesda, Maryland 20892, USA. [5]The Feinstein Institute for Medical Research, North Shore-Long Island Jewish Health System, Manhasset, New York 11030, USA. [6]Rheumatology Unit, Department of Medicine, Karolinska Institutet at Karolinska University Hospital Solna, Stockholm, Sweden. [7]Department of Rheumatology, Leiden University Medical Centre, Leiden, The Netherlands. [8]Celera, Alameda, California, USA. [9]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College, London. [10]Clinical Immunology and Rheumatology, Academic Medical Center, University of Amsterdam, The Netherlands. [11]Jan van Breemen Institute, The Netherlands. [12]Sanquin Research Landsteiner Laboratory, Academic Medical Center, University of Amsterdam, The Netherlands. [13]Laboratory of Immunogenetics, Department of Pathology and [14]Department of Rheumatology, VU University Medical Center, Amsterdam, The Netherlands. [15]Rosalind Russell Medical Research Center for Arthritis, Department of Medicine, University of California, San Francisco, California, USA. [16]University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. [17]Rowe Program in Genetics, University of California at Davis, Davis, California 95616, USA. [18]SeraCare Life Science, Cambridge, Massachusetts, USA. [19]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [20]Arthritis Research Campaign (arc)–Epidemiology Unit, Stopford Building, The University of Manchester, Manchester M13
Q2 9PT, UK. [21]Genome Institute of Singapore, Singapore. Correspondence should be addressed to R.M.P. (rplenge@partners.org).

Received 5 March; accepted 10 July; published online XX XX 2008; doi:10.1038/ngXXX

Table 1 Sample collections

		Control		Antibody			Genotyping	Case-control
	Case collection	collection	Origin	status	Cases	Controls	platform	stratification
Meta-analysis	North American Rheumatoid Arthritis Consortium (NARAC)	New York Cancer Project, New York City	North America	100% CCP+	873	1,196	Illumina 550K	Identity-by-state clustering
3,393 cases; 12,460 controls	Epidemiological Investigation of Rheumatoid Arthritis (EIRA)	EIRA	Sweden	100% CCP+	660	658	Illumina 317K	Epidemiologically matched, identity-by-state clustering
	Wellcome Trust Case Control Consortium (WTCCC)	Shared controls, multiple non-autoimmune diseases	United Kingdom	80% CCP+, 84% RF+	1,860	10,606	Affymetrix 500K	Geographically matched
Stage 1 replication	Nurses Health Study (NHS)	NHS	North America	100% RF+ or CCP+	257	411	Sequenom iPlex	Epidemiologically matched
1,089 cases; 1,862 controls	Brigham Rheumatoid Arthritis Sequential Study (BRASS)	National Institutes of Mental Health (NIMH)	Boston, USA	100% CCP+	407	814	Sequenom iPlex, Affymetrix 500K	Case-control matching with GWAS data
	NARAC II	New York Cancer Project, New York City	North America	100% CCP+	425	637	Sequenom iPlex	Case-control matching with ancestry informative markers
	NARAC III	Publicly available Shared controls	North America	100% CCP+	869	1,303	Illumina 317K	Case-control matching with GWAS data
Stage 2 replication	Genomics Collaborative Initiative (GCI)	GCI	North America	100% RF+	457	460	Kinetic PCR	Epidemiologically matched
2,840 cases; 3,945 controls	Leiden University Medical Center (LUMC)	LUMC	Leiden, The Netherlands	100% RF+ or CCP+	528	540	Kinetic PCR	Geographically matched
	EIRA-II	EIRA-II	Sweden	100% CCP+	435	412	Sequenom iPlex	Epidemiologically matched
	Genetics Network Rheumatology Amsterdam (GENRA)	GENRA	Amsterdam, The Netherlands	100% CCP+	551	1,230	Sequenom iPlex	Geographically matched

GWA data from three meta-analysis collections were used to identify candidate SNPs for replication. The replication set was divided into two stages: stage 1 replication (three collections) and stage 2 replication (five collections). For each collection we list the geographic origin, the source of the controls, the autoantibody status of cases, and the number of cases and controls. We also list the genotyping technology used to type SNPs of interest. Finally, we specify the strategy used to correct for case-control population stratification.

Wellcome Trust Case Control Consortium (WTCCC) with an Affymetrix 500K platform that (i) passed strict quality control criteria and (ii) were also present in the Phase II HapMap. We used the software package IMPUTE[12] to determine genotypes of these SNPs in individuals from Sweden (Epidemiological Investigation of Rheumatoid Arthritis, EIRA) and North America (North American Rheumatoid Arthritis Consortium, NARAC) on the basis of available Illumina platform SNP data (**Supplementary Fig. 1** online). To conduct a meta-analysis of SNP association with rheumatoid arthritis risk, we used the Cochran-Mantel-Haenszel (CMH) statistical test using genotype counts from the WTCCC and imputed probabilistic allele dosages in EIRA and NARAC. The CMH test allowed us to conduct a stratified analysis that maintained the three case-control collections as separate strata. CMH also allowed for further sub-stratification of EIRA and NARAC individuals into more homogenous subgroups using identity-by-state similarity for SNPs across the genome[2] to correct for residual population stratification. This resulted in improved genomic control inflation for both EIRA ($\lambda_{GC} = 1.03$) and NARAC ($\lambda_{GC} = 1.20$). As there was little evidence of population stratification in the WTCCC ($\lambda_{GC} = 1.06$), we did not further sub-stratify those individuals.

After calculating case-control CMH association statistics in the GWA meta-analysis, we observed minimal inflation for SNPs outside the MHC region ($\lambda_{GC} = 1.09$, $\lambda_{GC} = 1.02$ after normalizing to a 1,000 case and control collection, **Supplementary Fig. 2** online). Thus, there was little evidence of bias due to technical artifact or population stratification. In **Table 2** we present association statistics for validated and suggestive rheumatoid arthritis risk loci in European populations. Of the confirmed non-MHC risk loci, we observed association at *PTPN22*, 6q23/*TNFAIP3*, *STAT4* and Q7 *TRAF1-C5*. We also observed evidence of association at 4q27 (containing the *IL2* and *IL21* genes) and *CTLA4*, but not *PADI4*. The 4q27 result is an independent replication, suggesting that this is a true-positive association. *CTLA4* replicates in the WTCCC ($P = 0.026$), providing further support for the role of this locus in rheumatoid arthritis, as suggested by previous studies in EIRA and NARAC[10].

After excluding published risk loci (including those in the MHC region) and correcting for residual inflation by λ_{GC}, we found that 78 SNPs remained with possible associations at a $P < 10^{-4}$ threshold (**Supplementary Table 2 online**). These SNPs were grouped into 38 Q8 independent regions on the basis of pairwise linkage disequilibrium

Table 2 Meta-analysis results from regions previously associated with rheumatoid arthritis

SNP					EIRA		NARAC		WTCCC		Meta-analysis		Minor allele frequency		Meta-analysis odds ratio	Published
Prior study	Meta-analysis	r^2	Locus	Region	χ^2	P	χ^2	P	χ^2	P	χ^2	P	RA	Non-RA	(95% CI)	odds ratio
rs2240340	rs11203367	0.97	*PADI4*	1p36	3.6	0.058	0.4	0.55	0.2	0.64	0.3	0.59	0.42	0.42	1.02 (0.96–1.08)	1.10
rs2476601	rs6679677	1	*PTPN22*	1p13	13.1	0.00029	30.2	3.8×10^{-08}	151.8	7.1×10^{-35}	184.3	5.7×10^{-42}	0.17	0.10	1.79 (1.65–1.94)	1.75
rs7574865	rs11893432	0.74	*STAT4*	2q32	2.7	0.10	3.9	0.050	6.0	0.014	11.9	0.00055	0.21	0.18	1.14 (1.06–1.23)	1.32
rs3087243	rs3087243	n.a.	*CTLA4*	2q33	4.2	0.041	4.3	0.037	4.9	0.026	11.8	0.00060	0.40	0.44	0.90 (0.85–0.95)	0.88
rs6822844	rs4572894	0.47	*IL2, IL21*	4q27	1.8	0.18	4.8	0.029	1.8	0.18	6.4	0.011	0.27	0.29	0.92 (0.86–0.97)	0.72
*HLA-DRB1*04*	rs6457620	n.a.	*HLA-DRB1*	6p21	145.1	2.0×10^{-33}	321.6	6.4×10^{-72}	439.5	1.4×10^{-97}	846.8	3.6×10^{-186}	0.72	0.50	2.55 (2.40–2.71)	~3
rs6920220	rs6920220	n.a.	*OLIG3, TNFAIP3*	6q23	7.6	0.0059	7.8	0.0053	22.6	2.0×10^{-06}	36.5	1.5×10^{-09}	0.26	0.22	1.24 (1.16–1.32)	1.22
rs10499194	rs1167223	0.5	*OLIG3, TNFAIP3*	6q23	1.9	0.17	4.9	0.026	10.7	0.0011	17.0	0.000038	0.36	0.39	0.88 (0.83–0.93)	0.75
rs3761847	rs10118357	0.97	*TRAF1, C5*	9q33	7.9	0.0050	22.1	2.6×10^{-06}	0.0	0.96	9.3	0.0023	0.46	0.44	1.10 (1.04–1.16)	1.32

For each SNP, we identified the best proxy in our study and the calculated LD to that proxy on the basis of CEU HapMap (reported as r^2). We present the χ^2 score and two-tailed P value in EIRA, NARAC and WTCCC, and in the meta-analysis.

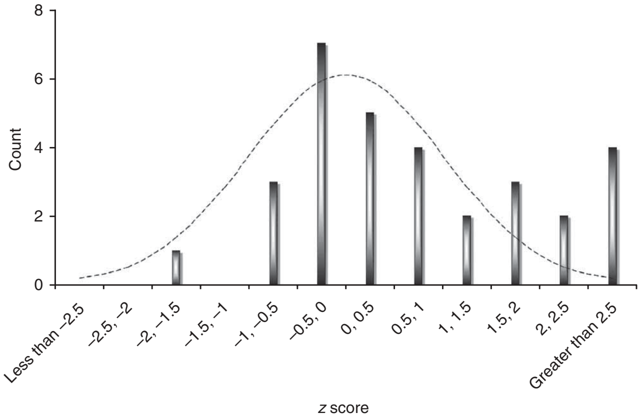

Figure 1 Enrichment of SNPs with z scores >2 in replication samples. For each of the 31 SNPs tested, we calculated a one-sided CMH z-score statistic from our two-staged replication data. Results were calculated using either stage 1 replication samples only (n = 14 SNPs) or using both stage 1 and 2 replication samples (n = 17 SNPs). A z score of 0 corresponds to a P value of 0.5, and a z score of 1.65 corresponds to a P value of 0.05. For a random collection of unassociated SNPs, this histogram should approximate a normal distribution (dotted line).

(LD) estimates derived from CEU HapMap (where SNPs were grouped together if $r^2 > 0.1$). We tested the single most significant SNP from each region in a two-staged replication.

Our replication collection consisted of eight independent case-control collections totaling 3,929 autoantibody (either CCP or RF) positive rheumatoid arthritis cases and 5,807 matched controls, all self-described as white and of European ancestry (**Table 1**). The presence of CCP or RF autoantibodies assures specificity for the diagnosis of rheumatoid arthritis and helps to minimize clinical heterogeneity across the eight collections. For each of the collections we further addressed potential case-control population stratification by either (i) using epidemiologically matched samples or (ii) matching cases and controls with ancestry informative genetic data; detailed strategies for each collection are described in the **Supplementary Note** online.

A total of 31 of these SNPs were successfully genotyped in all three stage 1 replication collections. The 17 most significant SNPs were genotyped in the stage 2 replication collections. For each SNP we calculated the replication P value using a one-tailed CMH statistic across the replication collections, and for those that replicated with $P < 0.05$, we calculated an overall P value (replication and the three meta-analysis collections) using a two-tailed CMH statistic (**Table 3**).

Testing in our complete replication set identified rheumatoid arthritis–associated alleles (**Table 2** and **Supplementary Tables 4** and **5** online). In replication genotyping, we observed that 6 out of 31 SNPs obtained $P \leq 0.01$; this is significantly more than expected by chance alone ($P = 9 \times 10^{-7}$ by Poisson). **Figure 1** illustrates the observed one-tailed CMH replication z scores, which clearly show that our results are enriched for $z > 2$ values (which corresponds to $P < 0.023$). Also, 4 of the 340,000 SNPs tested initially in the meta-analysis are associated with $P < 5 \times 10^{-7}$ in joint analysis; this is also significantly more than expected by chance alone ($P = 3 \times 10^{-5}$ by Poisson).

One SNP, rs4810485 in the 20q13 region, surpasses a conservative level of significance in joint analysis ($P = 8.2 \times 10^{-9}$) and thus represents a confirmed rheumatoid arthritis risk variant (**Fig. 2**). This SNP is located in the second intron of *CD40* and is within an LD block that contains a large portion of the *CD40* gene and its

Table 3 Newly identified SNPs associated with rheumatoid arthritis susceptibility

SNP						Meta-analysis				Replication				Joint		Breslow-Day test
				Allele		Frequency				Frequency						
ID	Chr.	Position	Locus	Major	Minor	Control	Case	p	OR	Control	Case	p	OR	p	OR	p
rs3890745	1p36	2585786	*MMEL1-TNFRSF14*	T	C	0.329	0.295	4.3×10^{-6}	0.86	0.321	0.301	0.0035	0.92	1.1×10^{-7}	0.89	0.31
rs42041	7q21	91891395	*CDK6*	C	G	0.243	0.274	5.5×10^{-5}	1.15	0.261	0.277	0.010	1.08	4.0×10^{-6}	1.11	0.43
rs2812378	9p13	34700260	*CCL21*	A	G	0.339	0.364	6.9×10^{-5}	1.13	0.334	0.355	0.00097	1.10	2.8×10^{-7}	1.12	0.67
rs4750316	10p15	6433266	*PRKCQ*	G	C	0.194	0.171	9.9×10^{-5}	0.86	0.197	0.183	0.0078	0.91	4.4×10^{-6}	0.88	0.77
rs1678542	12q13	56254982	*KIF5A-PIP4K2C*	C	G	0.373	0.352	5.4×10^{-6}	0.87	0.366	0.351	0.0026	0.92	8.8×10^{-8}	0.89	0.62
rs4810485	20q13	44181354	*CD40*	G	T	0.254	0.221	2.4×10^{-7}	0.83	0.246	0.231	0.0032	0.91	8.2×10^{-9}	0.87	0.37

We list results of 6 SNPs (out of 31 tested) that replicated with $P \leq 0.01$. The first six columns list SNP characteristics. The next four columns list GWA meta-analysis results including allele frequencies, a two-tailed P value for SNP association, and an odds ratio. The next four columns list similar results for replication genotyping; significance is reported on the basis of a stratified one-tailed CMH statistic. The next two columns summarize joint (overall) analysis results. Significance is reported on the basis of a stratified two-tailed CMH statistic across all 11 sample collections. The final column lists the Breslow-Day test for heterogeneity of odds ratios across all 11 collections.

5′ intergenic region. A SNP in near-perfect LD with rs4810485 ($r^2 = 0.95$, rs1883832) has been associated with autoimmune thyroid disease[13], although the association has not been confirmed unequivocally[14,15]. The same allele contributes to risk in both diseases. The rs1883832 variant has been shown to influence the efficiency of CD40 protein translation by disrupting a Kozak sequence[13]. The CD40 protein is expressed on the surface of multiple immune cells, including B cells, monocytes and dendritic cells, whereas its ligand, CD154, is expressed by activated CD4$^+$ T cells. CD40-CD154 interactions play a pivotal role in provision of helper activity by CD4$^+$ T cells in immune reactions including immunoglobulin class switching, memory B-cell development and germinal center formation[16]. Null mutations in *CD40* are known to cause a rare B cell–dependent hyper-IgM immunodeficiency syndrome[17].

The rs2812378 SNP in the 9p13 region replicates convincingly ($P = 0.00097$) and has a highly suggestive level of significance with an

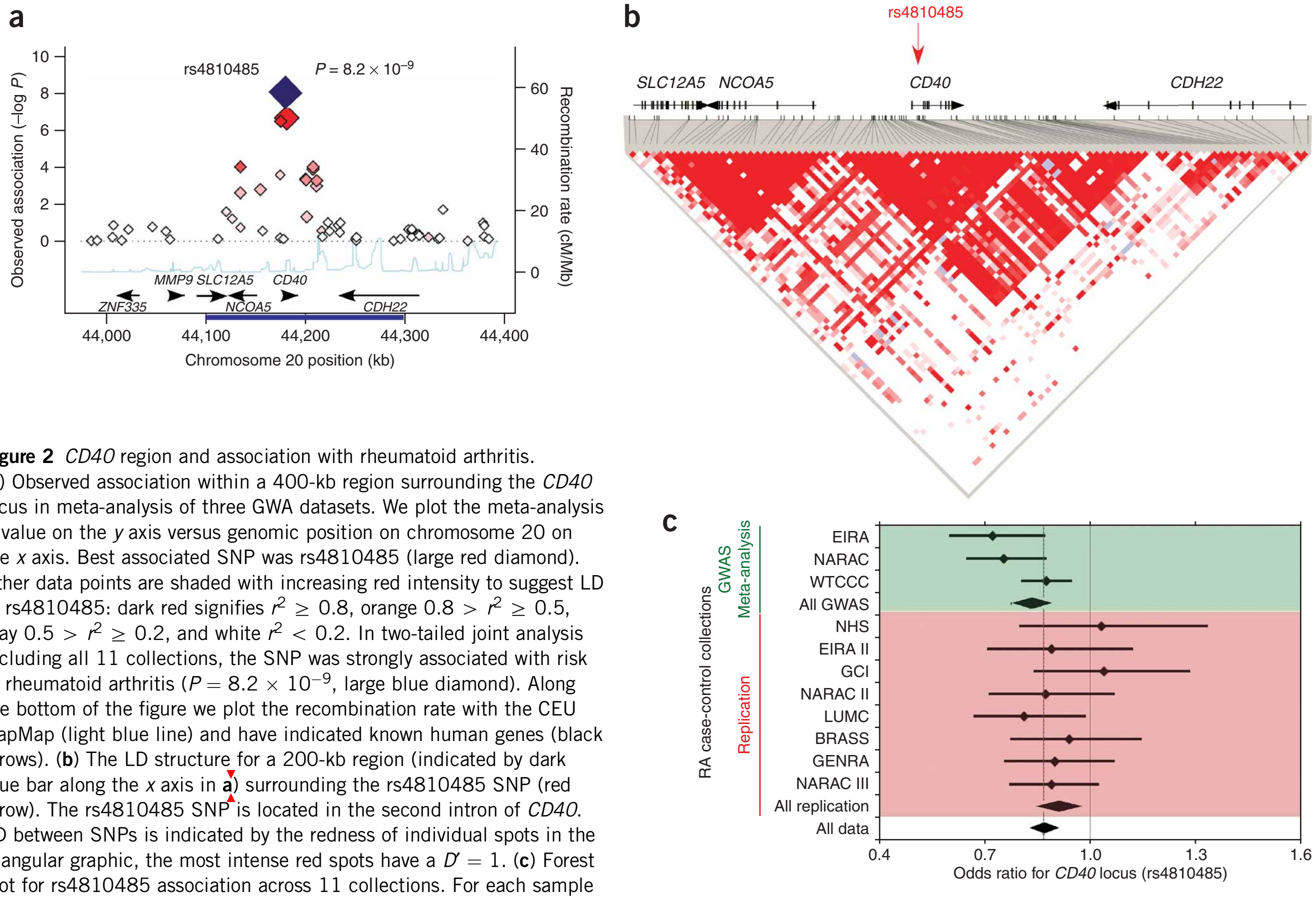

Figure 2 *CD40* region and association with rheumatoid arthritis. (**a**) Observed association within a 400-kb region surrounding the *CD40* locus in meta-analysis of three GWA datasets. We plot the meta-analysis P value on the y axis versus genomic position on chromosome 20 on the x axis. Best associated SNP was rs4810485 (large red diamond). Other data points are shaded with increasing red intensity to suggest LD to rs4810485: dark red signifies $r^2 \geq 0.8$, orange $0.8 > r^2 \geq 0.5$, gray $0.5 > r^2 \geq 0.2$, and white $r^2 < 0.2$. In two-tailed joint analysis including all 11 collections, the SNP was strongly associated with risk of rheumatoid arthritis ($P = 8.2 \times 10^{-9}$, large blue diamond). Along the bottom of the figure we plot the recombination rate with the CEU HapMap (light blue line) and have indicated known human genes (black arrows). (**b**) The LD structure for a 200-kb region (indicated by dark blue bar along the x axis in **a**) surrounding the rs4810485 SNP (red arrow). The rs4810485 SNP is located in the second intron of *CD40*. LD between SNPs is indicated by the redness of individual spots in the triangular graphic, the most intense red spots have a $D' = 1$. (**c**) Forest plot for rs4810485 association across 11 collections. For each sample collection we plot the odds ratio (small diamond) and the 95% CI. A light dashed line indicates the final odds ratio across all collections. The top three bars (green) represent GWA data and the diamond below them summarizes their meta-analyzed effect. The next eight bars (red) represent replication data and the diamond below them summarizes their meta-analyzed effect. The final diamond at the bottom of the page (clear) represents the final meta-analysis odds ratio and the 95% CI for all 11 collections.

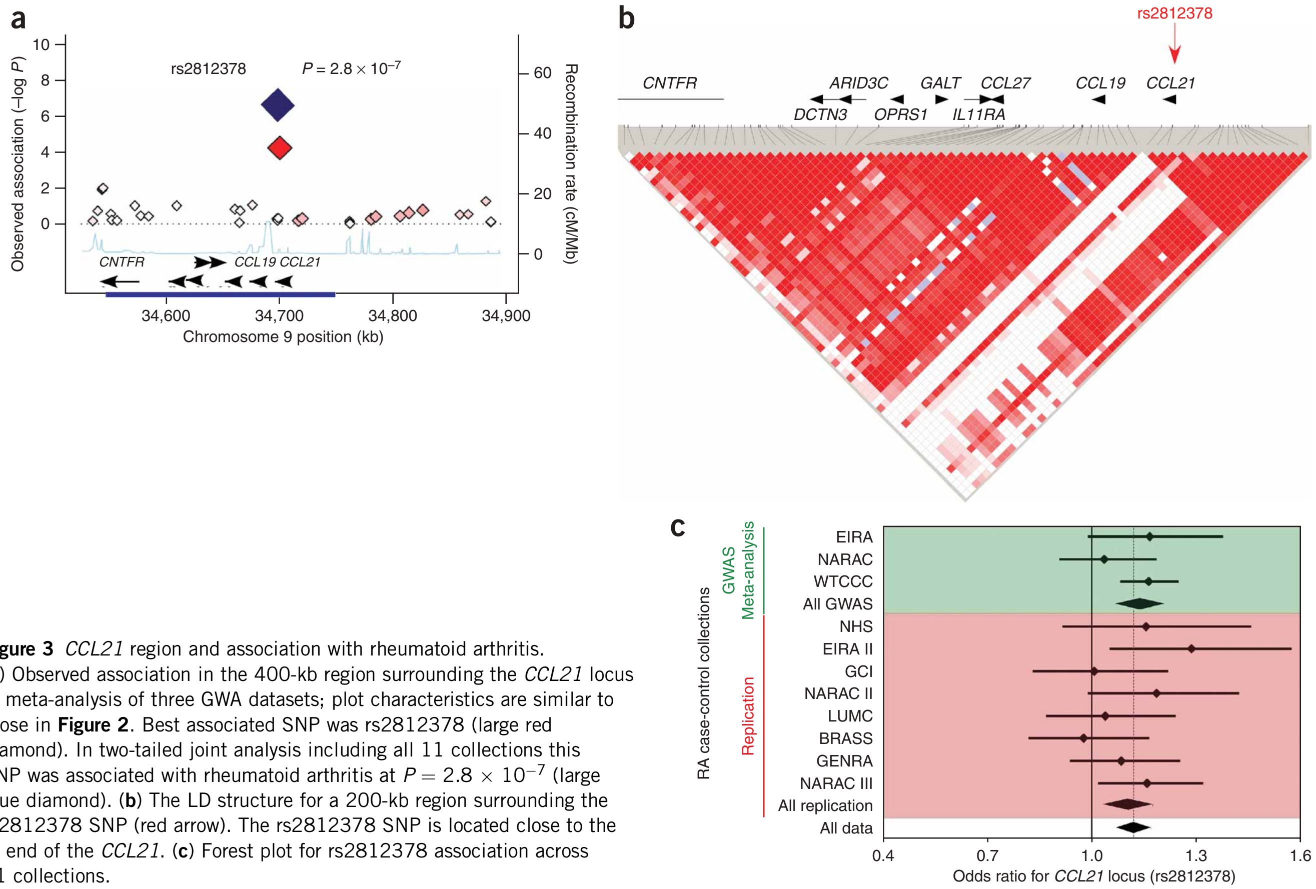

Figure 3 *CCL21* region and association with rheumatoid arthritis. (**a**) Observed association in the 400-kb region surrounding the *CCL21* locus in meta-analysis of three GWA datasets; plot characteristics are similar to those in **Figure 2**. Best associated SNP was rs2812378 (large red diamond). In two-tailed joint analysis including all 11 collections this SNP was associated with rheumatoid arthritis at $P = 2.8 \times 10^{-7}$ (large blue diamond). (**b**) The LD structure for a 200-kb region surrounding the rs2812378 SNP (red arrow). The rs2812378 SNP is located close to the 5′ end of the *CCL21*. (**c**) Forest plot for rs2812378 association across 11 collections.

overall $P = 2.8 \times 10^{-7}$ (**Fig. 3**). The SNP is located ~0.1 kb from the 5′ untranslated region of the *CCL21* gene, and is near a cluster of other genes including *CCL19* and *CCL27*. However, it is in an LD block that fully includes the *CCL21* coding sequence and not the other genes. The CCL21 protein is a chemokine that is involved in homing lymphocytes to secondary lymphoid organs. Expression of this chemokine is associated with ectopic lymphoid structures and has been implicated in the organization of lymphoid tissue affected by rheumatoid arthritis[18].

Most of the four other SNPs with $P \leq 0.01$ in our combined stage 1 and 2 genotyping probably represent true rheumatoid arthritis susceptibility alleles, although additional validation in large sample collections is required. Of the four regions, each contains genes that are known to be critical to the immune system. The rs42041 SNP on chromosome 7 maps to a *CDK6* intron; CDK6 is a ubiquitous cyclin-dependent kinase that regulates cell cycle progression, and it has been identified as a key mediator in the rapid proliferation of B cells and CD8 memory cells[19,20]. The rs1678542 SNP on chromosome 12 is ~20 kb away from the *PIP4K2C* locus, which has been implicated in signaling through the B-cell antigen receptor[21]. The rs3890745 SNP on chromosome 1 is ~60 kb away from *TNFRSF14*, which is similar to *CD40* in that it is a member of the TNF receptor super-family; it is known to bind *TRAF* family members including *TRAF1* and is involved in activation of the transcription factors NF-κB and AP-1 (ref. 22). The rs4750316 SNP on chromosome 10 is ~100 kb away from the 3′ end of the *PRKCQ* gene, which encodes a kinase required for the activation of the transcription factors NF-κB and AP-1, and may link the T-cell receptor signaling complex to the activation of the transcription factors[23].

A parallel UK study investigating the most significantly associated SNPs within the WTCCC study[1] has independently confirmed associations to *PRKCQ*, *KIF5A* and *MMEL1* in a separate cohort of 6,923 rheumatoid arthritis cases and an expanded reference group of 14,425 subjects[24].

Population stratification is unlikely to account for these observed effects, despite the modest effect sizes observed for rheumatoid arthritis risk ($0.87 \leq \text{OR} \leq 1.12$). We were careful to control for stratification individually in each of the meta-analysis GWA studies and also in each of the eight replication collections. Furthermore, the WTCCC collection contributed the greatest number of samples to the meta-analysis, and careful investigation across 12 subregions in the UK showed little evidence of case-control stratification[1]. Each of the associations presented here is notably significant in the WTCCC alone ($P < 0.001$). We found no evidence that different effects were present for these six loci across the five collections with genetically matched controls and the six collections with epidemiologically matched samples (Breslow-Day $P > 0.31$, **Table 2**). Technical artifact cannot explain the associations, as all SNPs passed strict quality control criteria (**Supplementary Table 3** online).

These associations provide strong evidence for the importance of the CD40 signaling pathway in autoantibody-positive rheumatoid arthritis. Our study implicates a putative functional variant that affects protein translation of the CD40 receptor. Established associations near *TRAF1* and *TNFAIP3* (also known as *A20*) already suggest the possibility that the CD40 signaling pathway mediates rheumatoid pathogenesis through NF-κB activation[25], although the rheumatoid arthritis risk variants have not yet been proven to modulate function of these genes. In particular, TRAF1 binds the CD40 receptor and Q11 cooperates with TRAF2 to activate NF-κB[26]. TRAF1 also binds TNFAIP3, which is a negative regulator of NF-κB signaling[27]. Furthermore, CD40 stimulation results in B-cell proliferation through regulation of *CDK6* expression[19]. The CD40 signaling pathway has been investigated in drug development, and mouse models have demonstrated that disruption of CD40 signaling could prevent development of immune-mediated arthritis and diabetes[28,29].

In conclusion, our study has identified an rheumatoid arthritis–associated variant for European populations at the *CD40* locus, provided strong evidence for association at the *CCL21* locus, and also suggests association at four other loci. It also provides empirical data suggesting that additional common alleles with odds ratios ~1.15 remain to be discovered. Even under the assumption that all of these variants are true risk factors, their total percent variance explained is only 0.6% (**Supplementary Note**). In this study, we estimate that the total percent variance explained for all known non-MHC common genetic variants is just 3.6%. Considering that ~60% of rheumatoid arthritis risk is thought to be genetic, and one-third of this risk is from the MHC locus[30], this indicates that less than half of genetic variation can be explained by the known rheumatoid arthritis risk alleles. One possibility is that there are other non-MHC common variants that have not yet been detected. All of the variants identified in our study have very modest effects and the rheumatoid arthritis Q12 case collections used in the meta-analysis were underpowered to screen for these effects at a modest level of significance ($P < 10^{-4}$). For example, assuming the observed odds ratios and allele frequencies, simulations show that the rs4810485 SNP in the *CD40* gene only had a 53% chance of meeting the $P < 10^{-4}$ significance criteria that we used to initially select SNPs. The other SNPs that replicate would have had only a 19–36% chance of being selected for further replication. Together, this suggests that other common alleles of modest effect size should be identified with additional GWA studies and deeper replication in large autoantibody-positive rheumatoid arthritis sample collections.

METHODS

Q13 **Subject groups.** Subject groups are described in detail in **Table 1** and in the **Supplementary Note**. Subjects were subdivided into three separate sets: (i) meta-analysis set, (ii) stage 1 replication set, and (iii) stage 2 replication set. Each collection consisted only of individuals that were self-described white and of European descent, and all cases either met 1987 ACR diagnostic criteria or were diagnosed by board-certified rheumatologists. Informed consent was obtained from each individual, and the institutional review board at each collecting site approved the study.

We used three subject groups to conduct the rheumatoid arthritis GWA meta-analysis. The groups used in the meta-analysis have been described elsewhere, and include those from the WTCCC, NARAC and EIRA. All of the cases in EIRA and NARAC and >80% of cases in WTCCC are CCP positive. For the WTCCC collection, we used an expanded collection of controls drawing from five non-autoimmune diseases that were genotyped as part of the larger study.

We used eight subject groups in our replication set. These collections consisted entirely of cases that were autoantibody positive (CCP or RF). For most of the collections, control samples were collected along with case samples as part of the same study. For some of the collections, control samples were unavailable; we matched these case collections to publicly available shared controls that had been genotyped on compatible platforms. The stage 1 replication set consisted of three subject groups: (i) CCP- or RF-positive cases identified by chart review from the Nurses Health Study (NHS) and matched controls based on age, gender, menopausal status, and hormone use; (ii) CCP-positive cases from the Brigham Rheumatoid Arthritis Sequential Study (BRASS) and controls from the National Institutes of Mental Health (NIMH); and (iii) CCP-positive cases drawn from North American clinics and controls from the New York Cancer Project (together this collection is called NARAC-II). The stage 2 replication set consisted of five collections: (i) CCP-positive cases drawn from North American clinics (NARAC-III) (P.K.G., unpublished data) and publicly available controls taken from a Parkinson's study and study 66 and 67 of the Illumina Genotype Control Database; (ii) North American RF-positive cases and controls matched on gender, age and grandparental country of origin from the Genomics Collaborative Initiative; (iii) CCP- or RF-positive Dutch cases and controls from Leiden University Medical Center (LUMC); (iv) CCP-positive cases from Sweden and epidemiologically matched controls (EIRA-II); and (v) CCP-positive Dutch cases and controls collected from the greater Amsterdam region (GENRA).

Genotyping. Detailed description of genotyping is provided in the **Supplementary Note**. All GWA meta-analysis genotyping was previously described. We directly genotyped 38 SNPs in stage 1 replication samples with the Sequenom iPlex platform at the Broad Institute (for NHS case-control samples and BRASS case samples) and National Institutes of Health (for NARAC-II and NYCP samples). We obtained NIMH genotypes from previously generated GWA data on the Affymetrix 500K platform through a formal application process. We genotyped stage 2 replication samples with (i) the Illumina 317K array at the Feinstein Institute (for the NARAC-III samples; unpublished data); (ii) using the kinetic PCR platform at Celera Diagnostics (for the GCI and LUMC samples); and (iii) with the Sequenom iPlex platform at the Broad Institute (for the EIRA-II and AMC/UVA samples). We obtained publicly available genotype data for shared controls for NARAC-III cases after an official application to a Parkinson's Disease consortium and Illumina Genotype Control Database. All stage 2 SNPs were directly genotyped in the GCI, LUMC, EIRA-II and GENRA samples, and individually imputed in NARAC-III case-control samples to determine genotype probability, as in our GWAS meta-analysis (see below).

In stages 1 and 2 we required that each SNP pass the following criteria for each collection separately: (i) genotype missing rate <10%, (ii) minor allele frequency >1%, and (iii) Hardy-Weinberg equilibrium with $P > 10^{-3}$. We also excluded individuals with data missing for >10% of SNPs. Of the 38 SNPs advanced into stage 1, 6 SNPs failed genotyping in Sequenom iPlex at either the Broad or NIH, and 2 failed in the NIMH dataset. The remaining SNPs had <4% missing data for each collection. All 17 SNPs passed stage 2 replication in NARAC-III, 2 failed in GCI and LUMC, and 4 failed in EIRA-II and GENRA.

Imputation and GWA meta-analysis. We conducted a GWA meta-analysis on a set of SNPs genotyped in the WTCCC study (**Supplementary Note**). We selected WTCCC SNPs on the basis of strict quality control criteria: (i) genotype missing rate <1% in cases and controls separately, (ii) minor allele frequency >1% in cases and controls separately, (iii) Hardy-Weinberg equilibrium with $P > 10^{-4}$ by a 2 degree-of-freedom test in cases and controls separately and (iv) availability of Phase II HapMap data. This resulted in a total of 336,721 SNPs. We imputed these SNPs in the EIRA and NARAC collections with IMPUTE. We used EIRA and NARAC data that had been filtered and imputed genotypes separately. We conducted separate runs for each chromosome using default parameters. As IMPUTE provides probabilistic confidence scores that track with prediction accuracy, we elected to use probabilistic dosages in our statistical analysis rather than hard genotype calls. This approach accounts for some uncertainty in imputation, and avoids bias.

To address case-control stratification we used identity-by-state to cluster EIRA cases and controls on the basis of on Illumina 317K SNP data into 165 substrata, and then to cluster NARAC cases and controls similarly on the basis of Illumina 550K SNP data into 396 substrata. This strategy was identical to that used to effectively control stratification previously in this dataset. As previous investigations revealed minimal case-control stratification in the WTCCC data, we placed all cases and controls from the WTCCC into a single stratum. We calculated association statistics using genotype counts available online for the WTCCC (see URLs section below) and probabilistic allele dosages for EIRA and NARAC. We calculated a CMH 1 degree-of-freedom test on the basis of allele frequency across 562 strata, and then after correcting I^2 scores by genomic control inflation (λ_{GC}), we assigned P values. Q14

Population stratification in replication samples. For each replication collection we corrected for possible case-control stratification by either (i) using only epidemiologically matched samples when cases and controls were drawn from the same population, or (ii) matching at least one control for each case on the basis of ancestry informative markers (see **Supplementary Note** for details). As the cases in the NHS, GCI, LUMC, EIRA-II, and GENRA collections were well matched to controls, we did not pursue further strategies to correct for case-control stratification. For the BRASS, NARAC-II and NARAC-III collections, we matched cases and controls with ancestry informative markers and placed them into a single stratum. For the BRASS cases and NIMH controls, GWA Q15 data on Affymetrix 6.0 (unpublished data) and Affymetrix 500K platforms were available, respectively. A total of 57,417 SNPs overlapped both datasets that had 0% missing data across all individuals; we used these as SNPs to derive ancestry information. For NARAC-II cases and NYCP controls, cases and controls were matched using genotype data on 760 ancestry informative markers. Finally for the NARAC-III cases (unpublished data) and shared controls, we used available Illumina 317K GWA data for 269,771 SNPs passing stringent quality control criteria. For each case-control collection, we used these SNPs to define the top ten principal components and to remove genetically distinct outliers (σ threshold = 6 with five iterations) with the software program EIGENSTRAT. We eliminated vectors that correlated with known structural variants on chromosomes 8 and 17, demonstrated minimal variation, or did not stratify cases and controls. After mapping cases and controls in the space of eigenvectors, we matched cases to controls that were nearest in Euclidean distance. A total of 814 of the available 1,498 NIMH controls were included (matching along the top two principal components), a total of 637 of the available 1,153 NARAC-II controls were included (matching along the top principal component), and a total of 1,303 of the available 2,189 NARAC-III shared controls Q16 were included (matching along the top two principal components).

Stage 1 and 2 replication analysis. We selected SNPs for replication by (i) identifying all SNPs that achieve $P < 10^{-4}$ significance in meta-analysis, (ii) grouping SNPs with $r^2 < 0.1$ into regions, and (iii) forwarding the SNP showing strongest association from each region for replication (**Supplementary Note**). We excluded SNPs from regions that had already demonstrated association in other studies. We forwarded 38 SNPs for stage 1 replication. The most significant SNPs from a preliminary statistical analysis conducted without correcting for possible case-control stratification were forwarded for stage 2 replication. For only those SNPs that replicated with $P < 0.05$, we Q17 genotyped EIRA samples and replaced imputed genotypes based on GWA data for our final analysis (**Supplementary Table 1** online).

For each SNP we conducted three statistical tests. First, we conducted a one-sided CMH statistical test across eight strata to assess whether rheumatoid arthritis association was reproducible in the replication collections in the same direction as the GWAS meta-analysis used to select the SNPs of interest. Second, we calculated a 570 strata joint analysis across all meta-analysis strata and substrata and replication strata; the eight replication collections were each placed into their own strata and the GWAS samples were partitioned into 562 strata, as described above. We considered $P < 5 \times 10^{-8}$ as a reproducible level of significance. Third, we calculated a Breslow-Day test of heterogeneity of odds ratios. We performed all analyses in MATLAB.

URLs. WTCCC genotype data, http://www.wtccc.org.uk/info/access_to_data_samples.shtml.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
We thank the WTCCC for making access to genotype data available online, and for providing autoantibody status of the WTCCC rheumatoid arthritis cases. We thank S. Myers and J. Marchini for help with IMPUTE. The BRASS Registry is supported by a grant from Millennium Pharmaceuticals and Biogen-Idec. S.R. is supported by a T32 NIH training grant (AR007530-23) and through the BWH Rheumatology Fellowship program, directed by S. Helfgott. R.M.P. is supported by a K08 grant from the NIH (AI55314-3), a private donation from the Fox Trot Fund, the William Randolph Hearst Fund of Harvard University, and holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund. M.J.D. is supported by a UO1 NIH grant (UO1 HG004171). The NARAC is supported by NIH grants RO1-AR44422 and NO1-AR-2-2263 (P.K.G.). This work was also supported in part by the Intramural Research Program of the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health. The EIRA study is supported by grants from the Swedish Medical Research Council, the Swedish Council for Working Life and Social Research, King Gustaf V's 80-year Foundation, the Swedish Rheumatic Foundation, the Stockholm County Council, the insurance company Arbetsmarknadens Försäkringsaktiebolag, and the County of Sörmland Research and Development Center. Genotyping of the GCI and LUMC samples was funded by Celera. D.A. is a Burroughs Wellcome Fund Clinical Scholar in Translational Research, and a Distinguished Clinical Scholar of the Doris Duke Charitable Foundation. B.D.K. was supported by the NIH Clinical Research Training Program, a public-private partnership between the Foundation for the National Institutes of Health and Pfizer Inc. L.A.C. is a Kirkland Scholar Awardee and her work on this project was supported by K24 AR02175, R01 AI065841 and 5 M01 RR-00079. P.P.T. and N.d.V. were supported by European Community's FP6 funding (Autocure). We acknowledge the help of C.Ellen van der Schoot for healthy control samples for GENRA and the help of B.A.C. Dijkmans, D. van Schaardenburg, A.S. Peña, P.L. Klarenbeek, Z. Zhang, M.T. Nurmohammed, W.F. Lems, R.R.J. van de Stadt, W.H. Bos, J. Ursum, M.G.M. Bartelds, D.M. Gerlag, M.G.H. van der Sande, C.A. Wijbrandts and M.M.J. Herenius in gathering GENRA subject samples and data. We thank Y. Li, S. Schrodi, and J. Sninsky (Celera); and B. Voight and C. Cotsapas (Broad Institute) for comments on the manuscript.

AUTHOR CONTRIBUTIONS
S.R., M.J.D. and R.M.P. designed the study, conducted the statistical analysis, interpreted the primary data, and wrote the initial manuscript. E.F.R. and B.D.K. generated Sequenom genotype data on the replication samples from North America (at NIAMS); N.P.B., R.H., L.G., S.W. and C.G. generated Sequenom genotype data on the replication samples from North America and NHS (at the Broad Institute); A.T.L. and P.K.G. generated the NARAC genome-wide association genotype data; A.B.B., M.C., K.G.A. and J.J.C. generated genotype data on GCI and LUMC samples (at Celera); F.A.S.K., R.E.M.T., A.H.M.M., and T.W.J.H. contributed LUMC samples; G.J.W., P.P.T., J.B.A.C., I.E.V.D.H.-B. and N.d.V. contributed GENRA samples; K.H.C., J.C. and E.W.K. contributed NHS samples and interpretation of the study data; L.K., L.P. and L.A. contributed EIRA samples and helped interpret the data; J.W. is principle investigator of the rheumatoid arthritis WTCCC study and contributed to the study design; N.A.S. and M.E.W. are principle investigators of BRASS, and J.C. contributed BRASS subject samples. P.K.G. is principle investigator of NARAC, provided replication samples and guidance on the study design, and helped to interpret the data. D.A. provided guidance on study design, interpretation of data and initial draft of manuscript. B.M.N. contributed statistical analysis. M.S. generated the genome-wide genotype data on EIRA. L.A.C., C.I.A., M.F.S., D.L.K., E.F.R., A.T.L., R.M.P. and P.K.G. are all members of NARAC and have contributed to the study design. All authors contributed to writing the final manuscript. Q22 Q18

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
2. Plenge, R.M. *et al.* TRAF1–C5 as a risk locus for rheumatoid arthritis–a genomewide study. *N. Engl. J. Med.* **357**, 1199–1209 (2007).
3. Kurreeman, F.A. *et al.* A candidate gene approach identifies the TRAF1/C5 region as a risk factor for rheumatoid arthritis. *PLoS Med.* **4**, e278 (2007).
4. Plenge, R.M. *et al.* Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat. Genet.* **39**, 1477–1482 (2007).
5. Thomson, W. *et al.* Rheumatoid arthritis association at 6q23. *Nat. Genet.* **39**, 1431–1433 (2007).
6. Stastny, P. Association of the B-cell alloantigen DRw4 with rheumatoid arthritis. *N. Engl. J. Med.* **298**, 869–871 (1978).
7. Begovich, A.B. *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.* **75**, 330–337 (2004).
8. Remmers, E.F. *et al.* STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N. Engl. J. Med.* **357**, 977–986 (2007).
9. Zhernakova, A. *et al.* 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**, 1284–1288 (2007).
10. Plenge, R.M. *et al.* 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**, 1044–1060 (2005).

11. Suzuki, A. *et al.* Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.* **34**, 395–402 (2003).
12. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
13. Jacobson, E.M. *et al.* A CD40 Kozak sequence polymorphism and susceptibility to antibody-mediated autoimmune conditions: the role of CD40 tissue-specific expression. *Genes Immun.* **8**, 205–214 (2007).
14. Heward, J.M. *et al.* A single nucleotide polymorphism in the CD40 gene on chromosome 20q (GD-2) provides no evidence for susceptibility to Graves' disease in UK Caucasians. *Clin. Endocrinol.* **61**, 269–272 (2004).
15. Houston, F.A. *et al.* Role of the CD40 locus in Graves' disease. *Thyroid* **14**, 506–509 (2004).
16. Kawabe, T. *et al.* The immune responses in CD40-deficient mice: impaired immunoglobulin class switching and germinal center formation. *Immunity* **1**, 167–178 (1994).
17. Lougaris, V., Badolato, R., Ferrari, S. & Plebani, A. Hyper immunoglobulin M syndrome due to CD40 deficiency: clinical, molecular, and immunological features. *Immunol. Rev.* **203**, 48–66 (2005).
18. Manzo, A. *et al.* Systematic microanatomical analysis of CXCL13 and CCL21 in situ production and progressive lymphoid organization in rheumatoid synovitis. *Eur. J. Immunol.* **35**, 1347–1359 (2005).
19. Ishida, T. *et al.* CD40 signaling-mediated induction of Bcl-XL, Cdk4, and Cdk6. Implication of their cooperation in selective B cell growth. *J. Immunol.* **155**, 5527–5535 (1995).
20. Veiga-Fernandes, H. & Rocha, B. High expression of active CDK6 in the cytoplasm of CD8 memory cells favors rapid division. *Nat. Immunol.* **5**, 31–37 (2004).
21. Carpenter, C.L. Btk-dependent of phosphoinositide synthesis. *Biochem. Soc. Trans.* **32**, 326–329 (2004).
22. Marsters, S.A. *et al.* Herpesvirus entry mediator, a member of the tumor necrosis factor receptor (TNFR) family, interacts with members of the TNFR-associated factor family and activates the transcription factors NF-kappaB and AP-1. *J. Biol. Chem.* **272**, 14029–14032 (1997).
23. Gruber, T. *et al.* PKCtheta cooperates with atypical PKCzeta and PKCiota in NF-kappaB transactivation of T lymphocytes. *Mol. Immunol.* **45**, 117–126 (2008).
24. Barton, A. *et al.* Identification of novel RA susceptibility loci at chromosomes 10p15, 12q13, and 22q13. *Nat. Genet.* advance online publication, doi:10.1038/ng.218 (14 September 2008).
25. Harnett, M.M. CD40: a growing cytoplasmic tale. *Sci. STKE* **2004**, pe25 (2004).
26. Xie, P., Hostager, B.S., Munroe, M.E., Moore, C.R. & Bishop, G.A. Cooperation between TNF receptor-associated factors 1 and 2 in CD40 signaling. *J. Immunol.* **176**, 5388–5400 (2006).
27. Song, H.Y., Rothe, M. & Goeddel, D.V. The tumor necrosis factor-inducible zinc finger protein A20 interacts with TRAF1/TRAF2 and inhibits NF-kappaB activation. *Proc. Natl. Acad. Sci. USA* **93**, 6721–6725 (1996).
28. Balasa, B. *et al.* CD40 ligand-CD40 interactions are necessary for the initiation of insulitis and diabetes in nonobese diabetic mice. *J. Immunol.* **159**, 4620–4627 (1997).
29. Durie, F.H. *et al.* Prevention of collagen-induced arthritis with an antibody to gp39, the ligand for CD40. *Science* **261**, 1328–1330 (1993).
30. MacGregor, A.J. *et al.* Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum.* **43**, 30–37 (2000).

QUERY FORM

NG	
Manuscript ID	[Art. Id: 233]
Author	
Editor	
Publisher	

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer queries by making the requisite corrections directly on the galley proof. It is also imperative that you include a typewritten list of all corrections and comments, as handwritten corrections sometimes cannot be read or are easily missed. Please verify receipt of proofs via e-mail

Query No.	***Nature of Query***
Q1	Please carefully check the spellings of all author names and affiliations.
Q2	Please provide postal codes for all affiliations.
Q3	OK as added?
Q4	Abstract OK as edited?
Q5	Please cite Table 3 within the text.
Q6	OK?
Q7	*TNFAIP3* IN 6q23? Please reword to avoid using a slash.
Q8	Please cite Supplementary Table 1 before Supplementary Table 2; all supplementary items must be called out in numerical order. If necessary, renumber the tables and resubmit the compiled SI file.
Q9	Please cite Supplementary Table 3 before Supplementary Tables 4 and 5.
Q10	OK as edited?
Q11	Previous sentence OK as edited?
Q12	OK as edited?
Q13	OK as edited?
Q14	Correct symbol?
Q15	By this group of authors, or others?
Q16	OK as edited?
Q17	Was replacement done ON THE BASIS OF GWA data, or were the imputed genotypes BASED ON GWA data?
Q18	Please spell out last names of "J.C."
Q19	OK to move here?
Q20	Correct?
Q21	Citation OK here?
Q22	Please provide information for Bo Ding.