nature genetics

REL, encoding a member of the NF-κB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis

Peter K Gregersen[1,14], Chistopher I Amos[2,14], Annette T Lee[1], Yue Lu[2], Elaine F Remmers[3], Daniel L Kastner[3], Michael F Seldin[4], Lindsey A Criswell[5], Robert M Plenge[6], V Michael Holers[7], Ted R Mikuls[8], Tuulikki Sokka[9], Larry W Moreland[10], S Louis Bridges, Jr[11], Gang Xie[12], Ann B Begovich[13] & Katherine A Siminovitch[12]

We conducted a genome-wide association study of rheumatoid arthritis in 2,418 cases and 4,504 controls from North America and identified an association at the *REL* locus, encoding c-Rel, on chromosome 2p13 (rs13031237, $P = 6.01 \times 10^{-10}$). Replication in independent case-control datasets comprising 2,604 cases and 2,882 controls confirmed this association, yielding an allelic OR = 1.25 ($P = 3.08 \times 10^{-14}$) for marker rs13031237 and an allelic OR = 1.21 ($P = 2.60 \times 10^{-11}$) for marker rs13017599 in the combined dataset. The combined dataset also provides definitive support for associations at both *CTLA4* (rs231735; OR = 0.85; $P = 6.25 \times 10^{-9}$) and *BLK* (rs2736340; OR = 1.19; $P = 5.69 \times 10^{-9}$). c-Rel is an NF-κB family member with distinct functional properties in hematopoietic cells, and its association with rheumatoid arthritis suggests disease pathways that involve other recently identified rheumatoid arthritis susceptibility genes including *CD40, TRAF1, TNFAIP3* and *PRKCQ*[1,2].

Rheumatoid arthritis (MIM180300) is a common autoimmune disorder affecting approximately 1% of individuals in populations of European origin and whose predominant manifestation is inflammation with bone and cartilage destruction in diarthrodial joints. The genetic basis for rheumatoid arthritis is complex, with at least six genes generally accepted as associated with disease in populations of European origin[1], including *HLA-DRB1*, *PTPN22*, *STAT4*, *TRAF1* and *TNFAIP3*. A number of additional loci have recently been reported as a result of expanded genome-wide association studies[3] and meta-analyses[2]. Many of these are likely to reflect true associations, although a convincing demonstration often requires very large samples sizes given that many of the associations at these loci are quite modest. In most cases, the causative allele(s) have not been identified, and therefore, the actual contribution to disease risk at these loci is unknown.

The diagnosis of rheumatoid arthritis is based on clinical criteria established over two decades ago[4]. However, these criteria do not yet include antibody reactivity to cyclic citrullinated peptides (CCP), the presence of which is a highly sensitive and specific marker for the diagnosis of rheumatoid arthritis; between 50% and 80% of individuals meeting standard criteria for rheumatoid arthritis have antibodies to CCP[5]. Notably, the classical *HLA-DRB1* associations with rheumatoid arthritis are entirely restricted to this phenotypic subgroup[6], as are many of the other reported genetic associations. In addition to this phenotypic heterogeneity, there is evidence for genetic heterogeneity in risk for rheumatoid arthritis among different ancestral groups[7], and this has complicated efforts at replication. Given these considerations, it is apparent that additional risk variants for rheumatoid arthritis remain to be discovered. For these reasons, we undertook an expansion of our previous genome-wide association study of rheumatoid arthritis[8], restricted to North American individuals of European origin that were overwhelmingly (~95%) positive for antibodies to CCP. We also assembled a large case-control population for replication studies.

All new genotyping of case samples for this study was conducted on Illumina HapMap370 BeadArray typing platforms, and after quality control filtering (see Online Methods), a combined dataset of 2,418 cases and 4,504 controls was available for genome-wide association analysis of 278,502 SNPs that passed all quality control filters applied to each set of data. The cases are derived in part from affected sibling pair families of the North American Rheumatoid Arthritis Consortium (NARAC) previously reported[8] (one case per

[1]The Feinstein Institute for Medical Research, North Shore-Long Island Jewish Health System, Manhasset, New York, USA. [2]University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. [3]Genetics and Genomics Branch, National Institute of Arthritis, Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, Maryland, USA. [4]Rowe Program in Genetics, University of California at Davis, Davis, California, USA. [5]Rosalind Russell Medical Research Center for Arthritis, Department of Medicine, University of California, San Francisco, California, USA. [6]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [7]University of Colorado Denver School of Medicine, Denver, Colorado, USA. [8]University of Nebraska Medical Center, Omaha, Nebraska, USA. [9]Jyväskylä Central Hospital, Jyväskylä, Finland. [10]University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA. [11]University of Alabama at Birmingham, Birmingham, Alabama, USA. [12]Mount Sinai Hospital and University Health Network, Toronto, Ontario, Canada. [13]Celera, Alameda, California, USA. [14]These authors contributed equally to this work. Correspondence should be addressed to P.K.G. (peterg@nshs.edu) or K.A.S. (ksimin@mshri.on.ca).

Received 5 February; accepted 1 May; published online 7 June 2009; doi:10.1038/ng.395

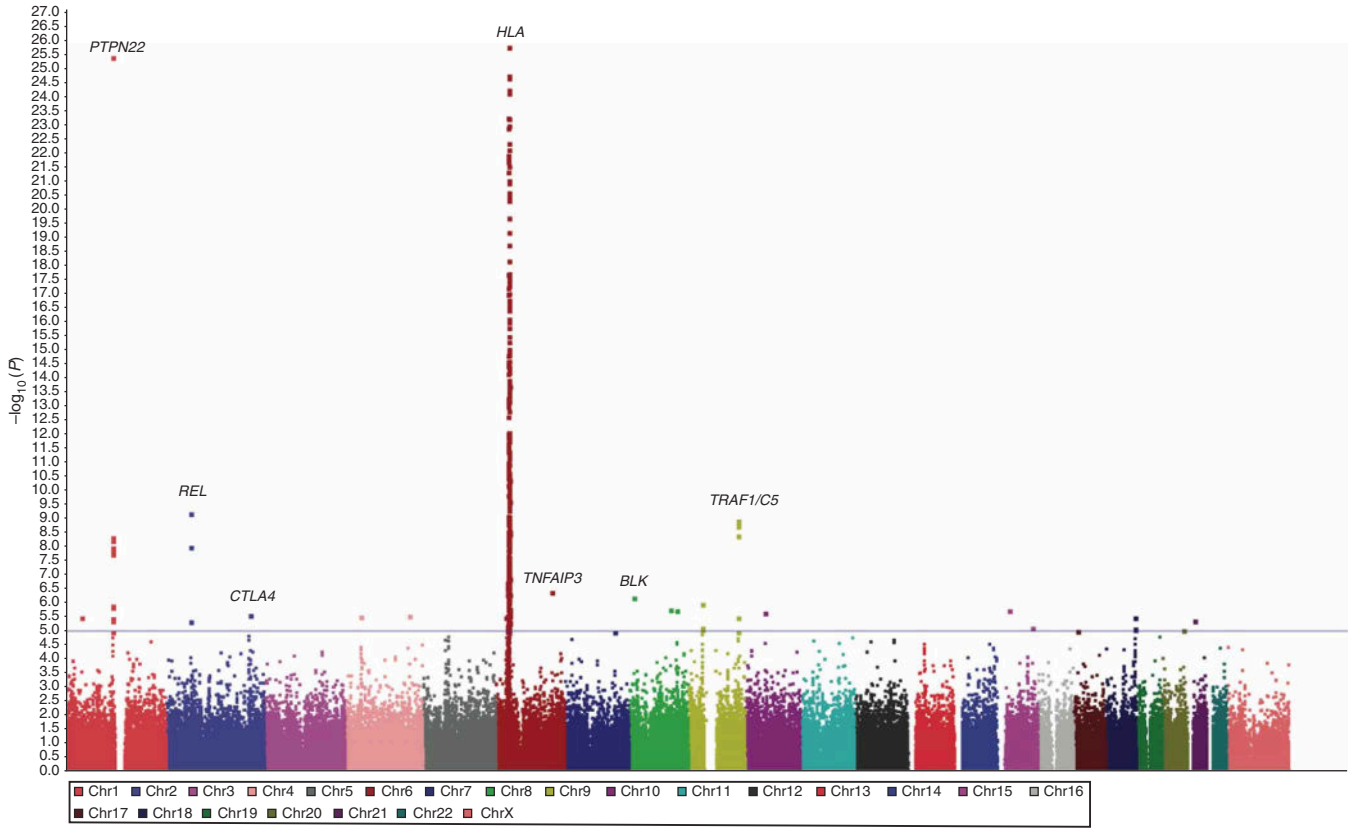

Figure 1 Summary of genome-wide association results for 2,418 cases and 4,504 controls. The $-\log_{10}$ values of the trend-test *P* values after homogeneous clustering and corrected for the residual genome-wide inflation of χ^2 are plotted against position on each chromosome. Chromosomes are shown in alternating colors for clarity. The blue horizontal line indicates SNPs that are significant at $P = 1 \times 10^{-5}$. The results for markers in the HLA region are truncated.

family), as well as new collections from both the United States and Canada (see **Supplementary Table 1** online). The genome-wide λ value inflation of χ^2 values was calculated to be 1.21, allowing for the large sample size. Structured association was therefore applied to correct for population stratification by matching homogeneous clusters of cases and controls, as implemented in PLINK. The genome-wide λ value after adjustment was calculated to be 1.06.

Figure 1 displays a graphical summary of the results of the genome-wide analysis as implemented in PLINK, after conditioning on clusters. As noted in previous studies, the largest association signal is in an extended region within the MHC achieving a maximal level significance of $P = 9.50 \times 10^{-104}$ in this study. (In **Fig. 1**, the *y* axis is truncated at –log $P = 27$.) In addition, we confirmed the previously reported associations[8] at *PTPN22* (rs2476601, $P = 1.62 \times 10^{-21}$) and the *TRAF1-C5* regions (maximal association at rs881375, $P = 4.09 \times 10^{-8}$). After *PTPN22*, the most significant association was observed on chromosome 2 in the region of *REL*, where two markers provide highly significant evidence of association (rs13031237, allele-specific $P = 6.01 \times 10^{-10}$ and rs13017599, allele-specific $P = 9.05 \times 10^{-9}$), as shown in **Table 1**. This is a new finding, as *REL* has not been brought forward as a candidate region for rheumatoid arthritis risk by any of the previous association studies in rheumatoid arthritis, although a recent publication has reported an association with celiac disease[9]. At lower levels of significance, we also noted evidence of association (**Table 1**) with *CTLA4* (rs6748358, $P = 8.24 \times 10^{-5}$) and a SNP marker in the region of the *BLK* locus (rs2736340, $P = 6.06 \times 10^{-7}$). A complete list of results at significance level $P < 0.01$ for the entire dataset is provided in **Supplementary Table 2** online.

To establish that these results are robust to various analytic approaches, we also conducted analysis using EIGENSTRAT. Before the principal components analysis (PCA), we removed markers in the region of chromosome 8p that show inversions in Northern Europeans (8.135–11.936 Mb) and markers in the centromeric region of chromosome 17q21.31 (40–43 Mb) that are polymorphic in populations of European origin[10]. We also removed markers around the HLA region that are related to European ancestry and also rheumatoid arthritis risk (from 24–36 Mb). These HLA-related markers were also removed for analyses that estimate the genome-wide inflation of test statistics that can arise from differences in population ancestry, as the HLA region includes many hundreds of markers that are associated with rheumatoid arthritis risk[6]. For the association tests with all markers (including those on chromosomes 6p, 8p and 17q), we adjusted for the eigenvectors derived using EIGENSTRAT to remove population admixture effects. The λ value was 1.20 before PCA correction and this value reduced to 1.06 after PCA correction. Results from EIGENSTRAT analysis along with trend tests from PLINK are presented in **Supplementary Table 3** online. Despite some genome-wide excess in the expected number of positive results from tests for association, the specific findings for *REL* did not seem to be influenced substantially by population structure. As shown in **Supplementary Table 3**, we found that, without and with adjustment for population structure, the *P* values and odds ratios for the SNPs we queried were quite similar. For example, for rs13031237, the

Table 1 Summary of association results for *REL*, *CTLA4* and *BLK*

Gene (alleles)	SNP	Study	Cases	Controls	Minor allele frequency		OR	95% CI	*P*
					Cases	Controls			
REL[a]	rs13031237	GWA	1,875/2,747	3,086/5,776	0.406	0.348	1.268	1.176–1.367	6.01×10^{-10}
(A/C)		Replication	1,536/2,494	1,579/3,085	0.381	0.339	1.207	1.104–1.320	3.55×10^{-5}
		Combined	3,411/5,241	4,665/8,861	0.394	0.345	1.246	1.177–1.318	3.08×10^{-14}
REL	rs13017599	GWA	1,837/2,751	3,071/5,773	0.400	0.347	1.248	1.157–1.345	9.05×10^{-9}
(A/G)		Replication	1,890/3,158	1,920/3,736	0.374	0.340	1.171	1.080–1.269	1.19×10^{-4}
		Combined	3,727/5,909	4,991/9,509	0.387	0.344	1.214	1.150–1.282	2.60×10^{-12}
CTLA4	rs231735	GWA	2,057/2,563	4,319/4,541	0.445	0.488	0.863	0.802–0.929	8.24×10^{-5}
(G/T)		Replication	2,154/2,750	2,706/2,930	0.439	0.480	0.869	0.804–0.939	3.81×10^{-4}
		Combined	4,211/5,313	7,025/7,471	0.442	0.485	0.855	0.812–0.902	6.25×10^{-9}
BLK	rs2736340	GWA	1,307/3,315	2,113/6,747	0.283	0.239	1.234	1.136–1.341	6.06×10^{-7}
(A/G)		Replication	1,338/3,732	1,372/4,276	0.264	0.243	1.122	1.027–1.225	1.08×10^{-2}
		Combined	2,645/7,047	3,485/11,023	0.273	0.240	1.194	1.125–1.268	5.69×10^{-9}

[a]GWA results: PLINK cmh test stratified by population-based identity-by-state clusters. Replication: cmh test stratified by five centers including Halifax, Toronto, OPA3, TEAR and GCI (except rs13031237 which was not genotyped in the GCI collection). Combined: cmh test stratified by eight centers, including NARAC I, NARAC II, Canada + lung cancer controls, Halifax, Toronto, OPA3, TEAR and GCI (except rs13031237 which was not genotyped in the GCI collection).

odds ratio and *P* value without adjustment for population structure in the combined genome-wide association analysis was 1.278 ($P = 5.2 \times 10^{-11}$) versus an odds ratio of 1.268 with adjustment ($P = 6.0 \times 10^{-10}$).

To confirm the associations with *REL*, we carried out a replication study on independent sets of 2,604 cases and 2,882 controls from the United States and Canada (**Supplementary Table 1b**). Although complete serologic data were not available for all subjects, most cases were seropositive (for either rheumatoid factor or CCP) in the replication datasets (see Online Methods). In addition to selected SNPs from the *REL* locus, we also included candidate SNPs from the *CTLA4* and *BLK* regions. For technical reasons, slightly different SNP panels were used for the replication studies in the Canadian and US samples on the basis of tagging linkage disequilibrium (LD) provided in HapMap; the results from the separate datasets are provided in **Supplementary Table 2**. The combined results with common SNP markers across all the datasets are shown in **Table 1** for *REL*, *CTLA4* and *BLK*. We observed two highly associated SNPs at the *REL* locus in the combined data using a Cochrane-Mantel-Haenszel analysis to allow for stratification among populations (rs13031237, OR = 1.24, $P = 3.08 \times 10^{-14}$ and rs13017599, OR = 1.21, $P = 2.06 \times 10^{-12}$). A graphical representation of the association results across the *REL* locus is shown in **Figure 2**. In addition, the data in **Table 1** (and **Supplementary Table 2**) provide definitive evidence for the previously suggested association with *CTLA4* (rs231735, OR = 0.86, $P = 6.25 \times 10^{-9}$). The data also support *BLK* as a newly identified rheumatoid arthritis risk locus (rs2736340, OR = 1.19, $P = 5.69 \times 10^{-9}$), a finding of some interest given the recent association of this locus with systemic lupus erethymatosus[11,12]. **Supplementary Table 4** online presents genotype-specific results, which show co-dominance for all the loci except *CTLA4*, which is nearly dominant.

The NF-κB (REL) family of transcription factors contains five members, including c-Rel, p65 (Rel-A), Rel-B, p50 (NFκB-1) and p52 (NFκB-2). These factors have a central role in coordinating the expression of a wide variety of genes that control immune responses and autoimmunity[13]. Therefore, the identification of *REL*, encoding c-Rel, as a new risk locus for rheumatoid arthritis has provoked us to consider how this observation may fit in with pathways suggested by the complex emerging landscape of genetic susceptibility for rheumatoid arthritis. Although the various NF-κB subunits have complex overlapping functions, current data suggest some distinct roles for c-Rel. The production of IL-12 and IL-23 subunits by macrophages and dendritic cells are critically dependent on c-Rel[14,15]. Thus, c-Rel knockout animals show deficiencies in Th1-type immune

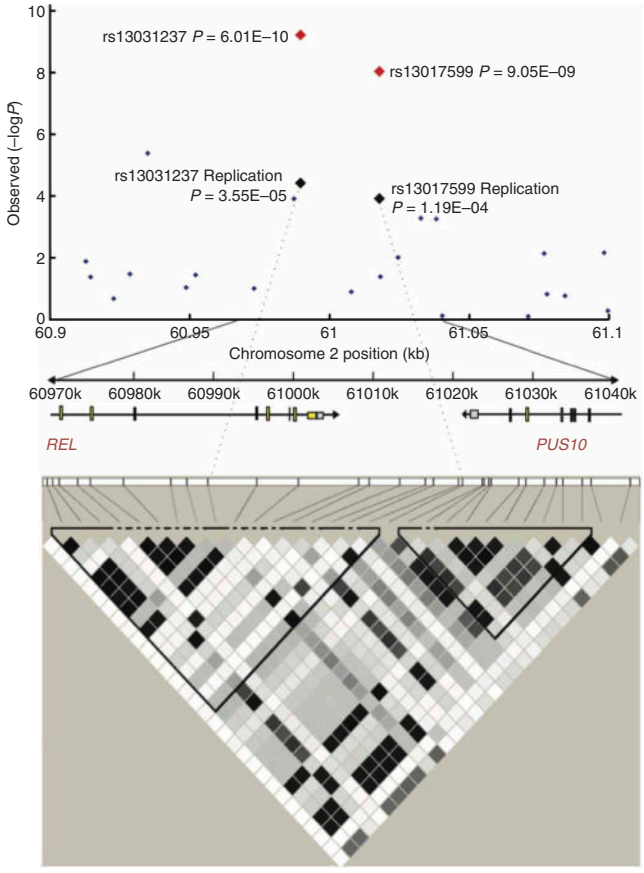

Figure 2 Association localization plots in the region around *REL* following discovery and replication phases. The *P* values for all samples in the GWA scan are shown as small blue diamonds, with the exception of two SNPs, rs13031237 and rs13017599 ($P = 9.05 \times 10^{-9}$ and $P = 6.01 \times 10^{-10}$) in *REL*, which are presented as red diamonds. The black diamonds show the independent results for these two SNPs in the replication samples.

responses, although intrinsic T-cell defects may also contribute to this phenotype[16]. Notably, both c-Rel and another gene recently identified to be involved in rheumatoid arthritis risk, *PRKCQ*, are specifically involved in the survival of activated CD8 cells, at least in part through the regulation of IL-2 production by these cells. In addition, a variety of genes in T cells are regulated by c-Rel, including *CD40* and *TNFAIP3*, both of which are now accepted rheumatoid arthritis susceptibility loci[17] and show evidence of association in this dataset (**Supplementary Table 5** online).

In addition to its effects on T-cell and antigen-presenting-cell function, c-Rel has been shown to have a role in B-cell proliferation and survival that particularly involves CD40 signaling pathways. Specifically, c-Rel–deficient B cells are susceptible to BCR-induced apoptosis that cannot be prevented by activation through CD40 (ref. 18), owing to reduced expression of the antiapoptotic protein Bcl-X_L, whose expression is known to be regulated by c-Rel. Notably, rescue from Fas-induced apoptosis is normal in these cells, demonstrating the existence of distinct CD40 signaling pathways that are at least in part distinguished by the involvement of c-Rel. A similar c-Rel–associated difference in CD40 signaling has been seen in individuals with ectodermal dysplasia and hyperIgM syndrome resulting from mutations in the NF-κB essential modulator, NEMO, where c-Rel–dependent IL-4 responses are also impaired[19]. c-Rel is the only NF-κB family member with oncogenic activity, and the gene is amplified in some B-cell lymphomas. In both tumors and normal B cells, it has recently been reported that the CD40 and c-Rel proteins can physically interact and form a heterodimer that is translocated to the nucleus[20] and regulates transcription of known c-Rel target genes, including *CD40LG* (*CD154*), *TNFSF13B* (*BLyS/BAFF*) and *BCL2A1* (*Bfl-1/A1*).

The associations of *CD40* (ref. 2), *REL*, *TRAF1* (ref. 8) and *TNFAIP3* (ref. 21) are consistent with an important role for CD40 signaling pathways in rheumatoid arthritis pathogenesis. Indeed, CD40–CD40L interactions have previously been identified as a potential target for therapy in autoimmune disease[22], and clinical trials in lupus have shown promise[23]. Unfortunately, the clinical development of monoclonal antibody inhibitors of CD40L was cut short by adverse effects on platelet function associated with the development of thromboembolic complications. Nevertheless, this pathway remains viable as a therapeutic target, and the current results mandate a thorough analysis of all the genes in this pathway to search for additional susceptibility alleles.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

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

ACKNOWLEDGMENTS

This work was supported by grants from the US National Institutes of Health NO1-AR-2-2263 (P.K.G.), RO1 AR44422 (P.K.G.) and by the Eileen Ludwig Greenland Center for Rheumatoid Arthritis and the Muriel Fusfeld Foundation. The work was also supported in part by the Intramural Research Program of the National Institute of Arthritis and Musculoskeletal and Skin Diseases and by grants from the Canadian Institutes for Health Research (MOP79321) and Ontario Research Fund (RE01061) and a Canada Research Chair to K.A.S.

AUTHOR CONTRIBUTIONS

P.K.G. organized and designed the study, supervised genotyping of US samples, contributed to statistical analysis and is the primary author of the manuscript. C.I.A. coordinated the design of the study, carried out and supervised all of the statistical analyses, and contributed to the writing of the manuscript. A.T.L. performed GWAS on all US samples and organized samples for distribution to carry out replication studies. E.L. carried out statistical analysis. E.F.R. contributed to study design, carried out replication genotyping and participated in preparation of the manuscript. D.L.K. contributed to study design and participated in review of the manuscript. M.F.S. contributed to study design and data analysis and participated in manuscript preparation. L.A.C. contributed to study design, contributed subjects for study and participated in manuscript preparation. R.M.P. contributed to study design and participated in manuscript preparation. V.M.H., T.R.M., T.S., S.L.B. and L.W.M. contributed subject populations for study and contributed to manuscript preparation. G.X. carried out genotyping and sample preparation of Canadian samples. A.B.B. contributed samples for study, carried out replication genotyping and participated in manuscript preparation. K.A.S. participated in study design and organization, supervised all genotyping of Canadian samples and participated in preparation of the manuscript.

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

1. Coenen, M.J. & Gregersen, P.K. Rheumatoid arthritis: a view of the current genetic landscape. *Genes Immun.* **10**, 101–111 (2009).
2. Raychaudhuri, S. *et al.* Common variants at *CD40* and other loci confer risk of rheumatoid arthritis. *Nat. Genet.* **40**, 1216–1223 (2008).
3. Barton, A. *et al.* Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat. Genet.* **40**, 1156–1159 (2008).
4. Arnett, F.C. *et al.* The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum.* **31**, 315–324 (1988).
5. Schellekens, G.A. *et al.* The diagnostic properties of rheumatoid arthritis antibodies recognizing a cyclic citrullinated peptide. *Arthritis Rheum.* **43**, 155–163 (2000).
6. Ding, B. *et al.* Different patterns of associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in the extended major histocompatibility complex region. *Arthritis Rheum.* **60**, 30–38 (2009).
7. Lee, H.S. *et al.* Genetic risk factors for rheumatoid arthritis differ in Caucasian and Korean populations. *Arthritis Rheum.* **60**, 364–371 (2009).
8. 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).
9. Trynka, G. *et al.* Coeliac disease associated risk variants in *TNFAIP3* and *REL* implicate altered NF-κB signalling. *Gut* advance online publication, doi:10.1136/gut.2008.169052 (24 February 2009).
10. Pennisi, E. Genetics. 17q21.31: not your average genomic address. *Science* **322**, 842–845 (2008).
11. Hom, G. *et al.* Association of systemic lupus erythematosus with *C8orf13-BLK* and *ITGAM-ITGAX*. *N. Engl. J. Med.* **358**, 900–909 (2008).
12. Harley, J.B. *et al.* Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM, PXK, KIAA1542* and other loci. *Nat. Genet.* **40**, 204–210 (2008).
13. Hayden, M.S., West, A.P. & Ghosh, S. NF-κB and the immune response. *Oncogene* **25**, 6758–6780 (2006).
14. Mise-Omata, S. *et al.* A proximal κB site in the IL-23 p19 promoter is responsible for RelA- and c-Rel-dependent transcription. *J. Immunol.* **179**, 6596–6603 (2007).
15. Grumont, R. *et al.* c-Rel regulates interleukin 12 p70 expression in CD8(+) dendritic cells by specifically inducing p35 gene transcription. *J. Exp. Med.* **194**, 1021–1032 (2001).
16. Mason, N.J., Liou, H.C. & Hunter, C.A. T cell-intrinsic expression of c-Rel regulates Th1 cell responses essential for resistance to *Toxoplasma gondii*. *J. Immunol.* **172**, 3704–3711 (2004).
17. Bunting, K. *et al.* Genome-wide analysis of gene expression in T cells to identify targets of the NF-κB transcription factor c-Rel. *J. Immunol.* **178**, 7097–7109 (2007).
18. Owyang, A.M. *et al.* c-Rel is required for the protection of B cells from antigen receptor-mediated, but not Fas-mediated, apoptosis. *J. Immunol.* **167**, 4948–4956 (2001).
19. Lu, K.T., Sinquett, F.L., Dryer, R.L., Song, C. & Covey, L.R. c-Rel plays a key role in deficient activation of B cells from a non-X-linked hyper-IgM patient. *Blood* **108**, 3769–3776 (2006).
20. Zhou, H.J. *et al.* Nuclear CD40 interacts with c-Rel and enhances proliferation in aggressive B-cell lymphoma. *Blood* **110**, 2121–2127 (2007).
21. Plenge, R.M. *et al.* Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat. Genet.* **39**, 1477–1482 (2007).
22. Toubi, E. & Shoenfeld, Y. The role of CD40–CD154 interactions in autoimmunity and the benefit of disrupting this pathway. *Autoimmunity* **37**, 457–464 (2004).
23. Boumpas, D.T. *et al.* A short course of BG9588 (anti-CD40 ligand antibody) improves serologic activity and decreases hematuria in patients with proliferative lupus glomerulonephritis. *Arthritis Rheum.* **48**, 719–727 (2003).

ONLINE METHODS

Subjects The cases and controls used in this analysis were taken from a variety of collections in both the United States and Canada, as detailed below. The genome-wide association data from 868 rheumatoid arthritis cases and 1,194 controls from North America have been previously reported in a study of susceptibility loci for rheumatoid arthritis[8]; all the additional genome-wide association data (1,550 cases and 3,310 controls) and all the replication sample data (2,604 cases and 2,882 controls) have not been previously reported in a rheumatoid arthritis association study. Informed consent was obtained from all subjects, and these studies were approved by the institutional review board of the North Shore LIJ Health System.

US samples. For our GWA analysis, we included GWA data on 868 North American rheumatoid arthritis cases samples reported previously[8]. We refer to these samples as 'NARAC I'. All cases included in the current analysis are positive for antibodies to CCP (CCP+). Details of sample collections used to compile the NARAC II dataset ($n = 952$) for final analysis on the Illumina HapMap370 chip are described below.

(1) Rheumatoid arthritis probands in NARAC collection of affected sibling pair families in whom siblings did not meet criteria for rheumatoid arthritis ($n = 17$), as well as probands of NARAC trio families ($n = 158$). Collection criteria have been previously published[8]. No affected members of these families were included in our previously published GWAS[8]. Only CCP+ subjects were studied in the current analysis.

(2) The Veterans Affairs Rheumatoid Arthritis Registry (VARA). VARA subjects, all meeting classification criteria for rheumatoid arthritis, are currently being enrolled from six VA centers (Dallas, Denver, Jackson, Omaha, Salt Lake City and Washington, DC). VARA is a longitudinal observational study that was started by T. Mikuls at the Omaha VA Medical Center. As of May 1, 2008, over 1,100 subjects have been enrolled. Of the 332 cases included in this analysis, 301 were documented to be CCP+ ($n = 332$).

(3) The Studies of Etiologies of Rheumatoid Arthritis (SERA) cohort. This cohort study, funded by the US National Institutes of Health (R01 AR051394), is co-directed by M. Holers and J. Norris at the University of Colorado Denver. SERA is a multicenter prospective cohort study designed to investigate genetic and epidemiologic associations with rheumatoid arthritis–related autoimmunity during the preclinical period of rheumatoid arthritis development. For the SERA study, probands with rheumatoid arthritis are recruited by rheumatologists from clinics at the University of Colorado Denver School of Medicine (UCDSOM), Cedars-Sinai Medical Center, Los Angeles, California, the Rheumatoid Arthritis Investigation Network (RAIN) centered in Omaha, Nebraska, North Shore-Long Island Jewish Health System/Feinstein Institute for Medical Research, and the Benaroya Research Institute at Virginia Mason, Seattle, Washington. Additionally, probands are recruited through community outreach efforts and advertisements posted in the Arthritis Foundation Newsletter. Once a proband with rheumatoid arthritis has expressed interest in the study and consents to release of medical information, their medical charts are reviewed by a SERA rheumatologist to ensure a diagnosis of rheumatoid arthritis. Probands are considered to have rheumatoid arthritis if they meet four of seven of the American College of Rheumatology (ACR) 1987 Revised Classification Criteria for rheumatoid arthritis or if they have been determined to have rheumatoid arthritis by a clinical evaluation by a board-certified rheumatologist. Only CCP+ cases were studied in the current analysis ($n = 160$).

(4) Multiple Autoimmune Disease Genetics Consortium (MADGC). The MADGC collection of multiplex families includes individuals with rheumatoid arthritis who meet 1987 ACR criteria for disease. The details of the MADGC collection have been previously published[24]. Only CCP+ cases were studied in the current analysis ($n = 105$).

(5) UCSF Rheumatoid Arthritis Genetics Project. Participants in the UCSF Rheumatoid Arthritis Genetics Project collection were recruited from UCSF Arthritis Clinics and private rheumatology practices in northern California as well as by nation-wide outreach according to a protocol approved by the University of California, San Francisco institutional review board. All cases fulfilled 1987 ACR criteria for rheumatoid arthritis and are of self-reported European origin, including grandparental countries of origin. Only CCP+ cases were studied in the current analysis ($n = 86$).

(6) Early Rheumatoid Arthritis Treatment Evaluation Registry (ERATER). This study in Nashville, Tennessee comprised 452 cases with rheumatoid arthritis of less than 3 years duration enrolled between February 2001 and August 2004, including 336 cases from a private practice of Arthritis Specialists of Nashville, 48 from Vanderbilt University Rheumatology Clinics and 68 from other sites. A detailed decription of this cohort has been published[25]. Only CCP+ cases were studied in the current analysis ($n = 94$).

Controls for the GWA were taken from control datasets that are publicly available in the Illumina iControl database. Additional control genotypes were from the neurodevelopmental control group obtained from the NIH Laboratory of Neurogenetics (see URLs section in Online Methods). These control genotypes were selected from the entire set of European American genotypes available from these resources on the basis of the following data filters: (i) $>$ 90% complete genotyping data, (ii) Hardy-Weinberg equilibrium $P > 10^{-4}$ and (iii) $>$90% European continental ancestry. The European continental ancestry was determined using ancestry informative markers as previously described[11]. In addition, we used controls derived from a recent GWA study of lung cancer[26]. None of these controls overlap with controls used for the previously reported NARAC GWAS[8]. A third set of 1,137 controls samples from M.D. Anderson Cancer Center Lung Cancer Study were used. These healthy individuals were seen for routine care at Kelsey-Seybold Clinics in the Houston Metropolitan area; all control subjects were required to have been current or former smokers.

The 'NARAC II Replication' dataset contained rheumatoid arthritis cases from the following sources:

(1) Treatment of Early Arthritis (TEAR) Trial. The TEAR trial is an investigator-initiated multicenter trial comparing two different strategies (early intensive therapy versus step-up therapy) and two combinations of medications (etanercept plus methotrexate (MTX) versus MTX plus hydroxychloroquine (HCQ) plus sulfasalazine (SSZ)). Subjects under study are those with early rheumatoid arthritis and an aggressive clinical phenotype defined by presence of active synovitis of multiple joints, with positive rheumatoid factor or baseline erosions as assessed by radiographs of hands or feet. We studied 347 rheumatoid factor–positive cases and 50 rheumatoid factor-negative cases ($n = 397$).

(2) The National Data Bank for Rheumatic Diseases (NDBRD). The National Databank is a longitudinal cohort of individuals with rheumatoid arthritis which has been described previously[27]. There is no overlap between the samples used in the previous GWAS and the samples used for replication reported here. Of the 312 cases included for study, 258 were documented to be CCP+, 15 were CCP–, and 39 subjects were not tested (total $n = 312$).

(3) The National Inception Cohort of Rheumatoid Arthritis (NICRA). The NICRA enrolled subjects within 6 months of clinical diagnosis and has been described in previous publications[28]. Only CCP+ cases were studied in the current analysis. There is no overlap between the samples used in the previous GWAS and the samples used for replication reported here ($n = 52$).

(4) Study of New Onset Rheumatoid Arthritis (SONORA). The SONORA enrolled individuals with rheumatoid arthritis or polyarthritis within 12 months after clinical diagnosis[29]. Only cases meeting 1987 ACR critieria for rheumatoid arthritis and who were CCP+ were included in the current study. There is no overlap between the samples used in the previous GWAS and the samples used for replication reported here ($n = 184$).

(5) The GCI (Genomics Collaborative, Inc.) sample set consisted of 475 rheumatoid arthritis cases and 475 individually matched controls. All case samples met the 1987 ACR diagnostic criteria for rheumatoid arthritis. All case samples were white North Americans of European descent who were positive for rheumatoid factor. Control samples were healthy individuals, also of European descent, with no medical history of rheumatoid arthritis. A single control was matched to each case on the basis of sex, age ($\pm$ 5 years) and self-reported ancestral background. Informed written consent was obtained from every subject.

For analytic purposes, the NDRBD, NICRA and SONORA cohorts were combined into a case group designated 'OPA3' as shown in **Supplementary Table 1b** and **Supplementary Table 2**.

Control subjects for NARAC II replication studies were taken in part from the New York Cancer Project[30] collection using subjects with self-reported European ancestry ($n = 1163$). Random matching of these controls (**Supplementary Table 1b**) with the TEAR and OPA3 replication datasets simply reflects the logistics of how replication genotyping was batched and has been

doi:10.1038/ng.395

used as a convenient way of grouping the analysis, shown in **Supplementary Table 2**. Additional controls were taken from the controls ($n = 474$) for the GCI collection as described above.

Canadian samples. Unrelated rheumatoid arthritis probands ($n = 598$) for the GWAS were recruited from the Toronto area; the diagnosis of rheumatoid arthritis was based on clinical, serological and radiological data in accordance with 1987 ACR criteria for rheumatoid arthritis. Subjects diagnosed with rheumatoid arthritis at an age of 16 years or younger were excluded from the study. Control samples ($n = 413$) were individuals of European origin from the Toronto area who had no history of rheumatoid arthritis or other inflammatory disease. Informed consent, demographic data and blood samples for genomic DNA extraction were obtained from every subject following institutional ethics committee approval.

A second, independent set of rheumatoid arthritis cases and healthy controls of European origin for replication was recruited from the Toronto ($n = 601$) and Halifax ($n = 564$) areas on the basis of the criteria described above. Controls for the replication study were recruited in Toronto ($n = 921$) and Halifax ($n = 324$), again with the criteria described above.

Genotyping and quality control filtering. The NARAC-II and Canadian case-control collections were genotyped using 373,400 SNPs on the Illumina HapMap370 BeadChip. Genotyping of the NARAC-II dataset was carried out at the Feinstein Institute for Medical Research, and the Canadian samples were genotyped at Illumina in San Diego, California. In order to organize the data to permit integration of results among studies, all genotypes were called using BeadStudio Software according to Top designation. Each dataset was subjected to quality control filtering based on SNP genotype call rates (>95% completeness), minor allele frequency (>0.01) and Hardy-Weinberg equilibrium ($P > 1 \times 10^{-4}$). We excluded subjects with more than 5% missing genotype data or showing evidence of non-European ancestry, as well as samples showing evidence of relatedness to another sample or possible DNA contamination. After filtering, we merged genotypes derived from SNP markers that were common to both datasets into a single file for analysis. A summary of all samples used for analysis after data cleaning is given in **Supplementary Table 1**. These filters were applied to each dataset independently and SNPs that passed quality control filters in each dataset were then merged. In all, a total of 278,502 SNPs passed all filters.

Replication genotyping of selected markers of interest was carried out using Sequenom iPlex technology at the Analytic Genetics Technology Centre in Toronto (Canadian samples) or within the Genetics and Genomics Branch of the National Institutes of Arthritis and Musculoskeletal and Skin Diseases (NARAC II Replication dataset). The GCI cohort was genotyped using a kinetic PCR assay at Celera. We selected slightly different panels of SNPs for replication in each dataset, on the basis of technical considerations for efficient multiplexing of markers. The patterns of linkage disequilibrium reported in the HapMap informed the selection of replacement markers to develop marker sets for the replication studies. For rs13031237 and rs2736340, the PCR reaction provided results on the alternate strand (T/G for A/C and T/C for A/G, respectively for rs13031237 and rs2736340).

To allow for potential effects of population substructure and heterogeneity among populations, we used several techniques. In initial analyses of genome-wide data, we combined marker genotypes of all individuals and conducted analyses to identify clusters of individuals who had similar genotypes using the program PLINK, with the clustering criterion PPC set at 0.0001. We also applied the program EIGENSTRAT to perform association analysis adjusting for correlations among the subjects according to marker similarities. For both clustering analyses in PLINK and construction of eigenvectors using EIGENSTRAT, we first removed SNPs on chromosome 6p near the HLA region and on regions in chromosomes 8p and 17q that contain large polymorphic inversions. We removed these SNPs as the HLA region contains many SNPs relating to case-control status and including them could reduce our power to detect true associations. The SNPs in chromosomes 8p and 17q show different LD patterns compared with other SNPs throughout the genome, which could influence the results owing to the potential chance variations caused by effects from many markers in these regions that are not of interest for correcting for population structure but that may show stronger correlations among subjects than expected.

To check for gene-gene interactions, we evaluated all possible interactions between SNPs in *CTLA4*, *REL* and *BLK* and the 157 most significantly associated non-HLA region SNPs with rheumatoid arthritis. Results showed no interaction effects with significance levels below $P < 0.001$. Given that 471 tests were performed, no significant interaction effects were identified after correcting for multiple testing.

λ values indicating population structure varied among the populations studied and according to the test conducted. For NARAC, the λ value after correcting to a sample of 1,000 cases and 1,000 controls using PLINK was 1.148 and for EIGENSTRAT λ was 1.029. For the Canadian sample, λ was 0.996 for PLINK and 1.010 for EIGENSTRAT, respectively. For the combined sample, the λ value was 1.077 for PLINK and 1.023 for EIGENSTRAT, respectively.

URLs. Illumina iControl database, www.illumina.com/iControlDB; NIH Laboratory of Neurogenetics, http://neurogenetics.nia.nih.gov/paperdata/public/; Top designation, http://www.illumina.com/downloads/TOPBOT_technote27Jun06.pdf.

24. Criswell, L.A. *et al.* Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes. *Am. J. Hum. Genet.* **76**, 561–571 (2005).
25. Sokka, T. & Pincus, T. An Early Rheumatoid Arthritis Treatment Evaluation Registry (ERATER) in the United States. *Clin. Exp. Rheumatol.* **23**, S178–S181 (2005).
26. Wang, Y. *et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat. Genet.* **40**, 1407–1409 (2008).
27. Wolfe, F., Michaud, K., Gefeller, O. & Choi, H.K. Predicting mortality in patients with rheumatoid arthritis. *Arthritis Rheum.* **48**, 1530–1542 (2003).
28. Fries, J.F. *et al.* HLA-DRB1 genotype associations in 793 white patients from a rheumatoid arthritis inception cohort: frequency, severity, and treatment bias. *Arthritis Rheum.* **46**, 2320–2329 (2002).
29. Weisman, M. *et al.* Analysis at one year of an inception cohort of early rheumatoid arthritis (RA): the SONORA study. *Arthritis Rheum.* **48**, S119 (2003).
30. Mitchell, M.K., Gregersen, P.K., Johnson, S., Parsons, R. & Vlahov, D. The New York Cancer Project: rationale, organization, design, and baseline characteristics. *J. Urban Health* **81**, 301–310 (2004).