Genetic variants at *CD28*, *PRDM1*, and *CD2/CD58* are associated with rheumatoid arthritis risk

Supplementary Note

Soumya Raychaudhuri[1-3 ††], Brian P. Thomson[2], Elaine F. Remmers[4], Stephen Eyre[5], Anne Hinks[5], Candace Guiducci[2], Joseph J. Catanese[6], Gang Xie[7], Eli A. Stahl[1], Robert Chen[1], Lars Alfredsson[8], Christopher I. Amos[9], Kristin G. Ardlie[2], BIRAC Consortium[†], Anne Barton[5], John Bowes[5], Noel P. Burtt[2], Monica Chang[6], Jonathan Coblyn[1], Karen H. Costenbader[1], Lindsey A. Criswell[10], J. Bart A. Crusius[11], Jing Cui[1], Phillip L. De Jager[2,12], Bo Ding[8], Paul Emery [13], Edward Flynn[5], Pille Harrison [14], Lynne J. Hocking[15], Tom W. J. Huizinga[16], Daniel L. Kastner[4], Xiayi Ke[5], Fina A. S. Kurreeman[1,16], Annette T. Lee[17], Xiangdong Liu[7], Yonghong Li[6], Paul Martin[5], Ann W. Morgan[13], Leonid Padyukov[18], David M. Reid[15], Mark Seielstad[19], Michael F. Seldin[20], Nancy A. Shadick[1], Sophia Steer [21], Paul P. Tak[22], Wendy Thomson[5], Annette H. M. van der Helm-van Mil[16], Irene E. van der Horst-Bruinsma[23], Michael E. Weinblatt[1], Anthony G. Wilson[24], Gert Jan Wolbink[25,26], Paul Wordsworth[14], YEAR Consortium[†], David Altshuler[2,3], Elizabeth W. Karlson[1], Rene E. M. Toes[16], Niek de Vries[22], Ann B. Begovich[6,27], Katherine A. Siminovitch[7], Jane Worthington[5], Lars Klareskog[18], Peter K. Gregersen[17], Mark J. Daly[2,3], Robert M. Plenge[1,2 ††]

[†]Supplementary Note Online
[††] Corresponding authors

1. Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, 02115, USA.
2. Broad Institute, Cambridge, Massachusetts, 02142 USA
3. Center for Human Genetic Research, 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. Arthritis Research Campaign (arc)–Epidemiology Unit, Stopford Building, The University of Manchester, Manchester M13 9PT, United Kingdom.
6. Celera, Alameda, California 94502, USA.
7. Dept of Medicine, University of Toronto, Mount Sinai Hospital and University Health Network, Toronto, Ontario M5G 1X5, Canada.
8. Institute of Environmental Medicine, Karolinska Institutet, Stockholm 171 77, Sweden.
9. University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, USA.
10. Rosalind Russell Medical Research Center for Arthritis, Department of Medicine, University of California, San Francisco, California 94143, USA.
11. Laboratory of Immunogenetics, Department of Pathology, VU University Medical Center, 1007 MB Amsterdam, The Netherlands.
12. Department of Neurology, Center for Neurologic Diseases, Brigham and Women's Hospital, Boston, MA 02115, USA.
13. NIHR-Leeds Musculoskeletal Biomedical Research Unit, Leeds Institute of Molecular Medicine, University of Leeds, LS9 7TF, United Kingdom.
14. University of Oxford Institute of Musculoskeletal Sciences, Botnar Research Centre, Oxford OX3 7LD, United Kingdom.
15. Musculoskeletal and Genetics Section, Division of Applied Medicine, University of Aberdeen, AB25 2ZD, United Kingdom.

16. Department of Rheumatology, Leiden University Medical Centre, 2333 ZA Leiden, The Netherlands.
17. The Feinstein Institute for Medical Research, North Shore-Long Island Jewish Health System, Manhasset, New York 11030, USA.
18. Rheumatology Unit, Department of Medicine, Karolinska Institutet at Karolinska University Hospital Solna, Stockholm 171 76, Sweden.
19. Genome Institute of Singapore, Singapore 138672.
20. Rowe Program in Genetics, University of California at Davis, Davis, California 95616, USA.
21. Clinical and Academic Rheumatology, Kings College Hospital NHS Foundation Trust, Denmark Hill, London SE5 9RS, United Kingdom.
22. Clinical Immunology and Rheumatology, Academic Medical Center, University of Amsterdam, Amsterdam 1105AZ, The Netherlands.
23. Department of Rheumatology, VU University Medical Center, 1007 MB Amsterdam, The Netherlands.
24. School of Medicine & Biomedical Sciences, Sheffield University, Sheffield S10 2JF, United Kingdom.
25. Jan van Breemen Institute, 1056 AB Amsterdam, The Netherlands.
26. Sanquin Research Landsteiner Laboratory, Academic Medical Center, University of Amsterdam, 1006 AD Amsterdam, The Netherlands.
27. Roche Diagnostics, Pleasanton, CA 94588 USA.

Yorkshire Early Arthritis Register (YEAR) consortium members are:

Management Team: Professor Paul Emery[1], Professor Philip Conaghan[1], Professor Ann Morgan[1], Dr. Mark Quinn[2], Dr. Anne-Maree Keenan[1], Dr. Elizabeth Hensor[1]

Consultants: Dr. Andrew Gough[3], Dr. Michael Green[2,3], Dr. Richard Reece[4], Dr. Lesley Hordon[5], Dr. Philip Helliwell1[6], Dr. Richard Melsom[6], Dr. Sheelagh Doherty[7], Dr. Ade Adebajo[8], Dr. Andrew Harvey[9], Dr. Steve Jarrett[9], Dr. Gareth Huson[1], Dr. Amanda Isdale[2], Dr. Mike Martin[1], Dr. Zunaid Karim[9], Professor Dennis McGonagle[10], Dr. Colin Pease[1], Dr. Sally Cox[1]

SpRs: Dr. Victoria Bejarano[1], Dr. Jackie Nam[1]

Nurses: Claire Brown[1], Christine Thomas[1], David Pickles[1], Alison Hammond[1], Beverley Neville[3], Alan Fairclough[4], Caroline Nunns[4], Anne Gill[2], Julie Green[2], Belinda Rhys-Evans[1], Barbara Padwell[1], Julie Madden[10], Lynda Taylor[10], Sally Smith[1], Heather King[1], Jill Firth[6], Jayne Heard[7], Linda Sigsworth[6]

Lab Staff: Diane Corscadden[1], Karen Henshaw[1], Lubna-Haroon Rashid[1], Stephen G Martin[1], James I Robinson[1]

[1]Section of Musculoskeletal Disease, LIMM, Leeds, UK
[2]York District Hospital, York, UK
[3]Harrogate District Hospital, Harrogate, UK
[4]Huddersfield Royal Infirmary, Huddersfield, UK
[5]Dewsbury District and General Hospital, Dewsbury, UK
[6]St Luke's Hospital, Bradford, UK
[7]Hull Royal Infirmary, Hull, UK
[8]Barnsley District General Hospital, Barnsley, UK
[9]Pinderfields General Hospital, Wakefield, UK
[10]Calderdale Royal Hospital, Halifax, UK

Biologics in RA Control (BIRAC) consortium members are:

Manchester Royal Infirmary - Central Manchester University Hospitals NHS Foundation Trust (Dr. I N Bruce, Dr. K Hyrich, Dr. P Ho, Dr. R Gorodkin)

Derby Hospital - Derby Hospitals NHS Foundation Trust (Dr. L J Badcock, Dr. C M Deighton, Dr. S C O'Reilly, Dr. N Raj, Dr. M R Regan, Dr.G D Summers, Dr. R A Williams)

Macclesfield Hospital - East Cheshire NHS Trust (Dr. A Barton, Dr. S M Knight, Prof. D P M Symmons)

Withington Hospital - Manchester Primary Care Trust (Dr. P Sanders, Dr. B Pal)

Norfolk and Norwich Hospital - Norfolk&Norwich University Hospital NHS Foundation Trust (Dr. K Gaffney, Prof. A J Macgregor, Dr. T Marshall, Dr. P Merry, Prof. D G I Scott)

Gartnavel Hospital - North Glasgow University Hospitals Division, Greater Glasgow NHS Trust (Dr. D Porter, Dr. R Madok, Dr. M Gordon, Prof. I McInnes, Prof. R Sturrock, Dr. M Field)

North Manchester General Hospital - Pennine Acute Hospitals NHS Trust (Dr. B Harrison, Dr. M Pattrick, Dr. H N Snowden, Dr. A P Bowden, Dr. E E Smith, Dr. P Klimiuk, Dr. D J Speden)

Poole Hospital NHS Foundation Trust (Dr. P W Thompson, Dr. S C M Richards)

Queen Aexandra Hospital - Portsmouth Hospitals NHS Trust (Dr. R G Hull, Dr. J M Ledingham, Dr. F Mccrae, Dr. M R Shaban, Dr. A L Thomas, Dr. S A Young Min)

Hope Hospital - Salford Royal NHS Foundation Trust (Dr. A L Herrick, Dr. R G Cooper, Dr. R Benitha, Prof. A K P Jones, Dr. T W O'Neill)

St Helens and Knowsley Hospitals NHS Trust (Dr. V E Abernethy, Dr. A R Clewes, Dr. J K Dawson, Dr. M Lynch)

Russells Hall Hospital - The Dudley Group of Hospitals NHS Foundation Trust (Prof. G Kitas, Dr. J P Delamere, Dr. N Erb, Dr. R Klocke, Dr. A J Whallett)

Haywood Hospital, Stoke - University Hospital of North Staffordshire NHS Trust (Dr. P T Dawes, Dr. C M Dowson, Dr. A Hassell, Prof. E M Hay, Dr. S Kamath, Dr. J Packham, Dr. R S Sandhu, Dr. M F Shadforth)

Chapel Allerton Hospital - Leeds Teaching Hospitals NHS Trust (Dr. S. Bingham, Prof. P Emery, Dr. A. Morgan, Prof. H A Bird, Prof. P G Conaghan, Dr. C T Pease, Dr. R J Wakefield)

Cannock Chase Hospital - Mid-Staffordshire NHS Foundation Trust (Dr. S V Chalam, Dr. D Mulherin, Dr. T Price, Dr. T Sheeran, Dr. S Venkatachalam)

Freeman Hosptial - The Newcastle upon Tyne Hospitals NHS Foundation Trust (Mr P Crook, Dr. H E Foster, Dr. B Griffiths, Dr. I D Griffiths, Dr. M L Grove, Prof. J D Isaacs, Dr. L Kay, Dr. W F Ng, Dr. A Myers, Dr. P N Platt, Dr. D J Walker)

Kings Mill Hospital - Sherwood Forest Hospitals NHS Foundation Trust (Dr. K L Lim, Dr. D A Walsh, Dr. N D Carter)

Rotherham District Hospital - The Rotherham NHS Foundation Trust (Dr. M Holt, Dr. F W Fawthrop, Dr. G W Smith)

A. Selecting SNPs for validation. As a starting point we used results from a meta-analysis of three genome wide association studies (GWAS) for rheumatoid arthritis. Nominally significant SNPs were tested for functional connections to established RA risk SNPs with GRAIL. The most connected subset was forwarded for additonal genotyping.

a. **Meta-analysis of three GWAS.** We had previously conducted a large scale meta-analysis of three GWAS, described in detail elsewhere[1]. Briefly, in that study we compiled data from three GWAS consisting in aggregate of 3,393 cases and 12,462 controls (see **Supplementary Table 3**). We used IMPUTE to interpolate missing SNP data to facilitate cross-platform analysis[2,3]. In that study we examined 336,721 sites outside the MHC region that passed strict quality control criteria. In the originial study, we specifically conducted followup genotyping on independent SNPs that demonstrated nominal association to RA at $p < 0.0001$ in a two stage replication experiment. In that study we genotyped 17 loci in both the first and second replication stage - of those, 6 demonstrated compelling evidence of association.

b. **Selecting Candidate SNPs for followup.** For this study we identified all of those SNPs that were nominally associated with RA ($p < 0.001$) - a total of 510 SNPs exceed this threshold of significance. We grouped SNPs into independent loci; we considered two SNPs as in the same locus if there was evidence of LD ($r^2>0.1$ in CEU HapMap). We removed all loci that overlapped with validated RA risk regions (see **Table 1** main text). We also removed 17 additional loci with $p < 10^{-4}$ that were genotyped in both stages of the initial meta-analysis study. We were left with a total of 179

independent loci representing 370 SNPs for followup. From these loci, we selected the single SNP that demonstrated the greatest evidence of association in the published meta-analysis for followup genotyping.

c. **Prioritizing Candidate SNPs with GRAIL.** Using the GRAIL algorithm[4] we tested these 179 SNPs against 16 independent known associated RA risk loci. The implemnetation of GRAIL that we used contained literature only up until December 2006, and contained information about 25,455 genes derived from 259,659 PubMed Abstracts; there was a median of 13 references per gene. GRAIL references were based on gene references listed in Entrez for both the human gene of interest, and also references listed for homologous genes in model organisms[5]. Of note - the version of GRAIL that we used is a previous implementation that differs slightly from the current implementations - results are not substantially affected when we do the same experiment with the current version of GRAIL (see **Supplementary Figure 2**). We entered the 179 SNPs as 'query' loci and tested them against the 16 known validated RA SNPs, which we entered as 'seed' loci. We selected the 22 SNPs with GRAIL scores that were $p_{text} < 0.01$. Analysis of validated RA SNPs suggested that this was a reasonable cutoff (see main text).

B. Independent Patient Collections for replication. The patient collections that we used for validation genotyping are described in detail in **Supplementary Table 3**. All cases and controls were self-described "white" and of European ancestry. All cases either fulfilled 1987 ACR[6] criteria or were diagnosed by a board certified rheumatologist. All cases were seropositive for either rheumatoid factor (RF) or anti-cyclic citrullinated peptide antibody (CCP)[7].

a. The Brigham Rheumatoid Arthritis Sequential Study (BRASS) is a registry of RA patients from the Boston area followed at Brigham and Women's Hospital; it is described in detail elsewhere[8]. A board certified rheumatologist at Brigham and Women's Hospital diagnosed each patient. We identified all patients that were CCP positive. We obtained healthy controls for these samples from three separate studies on Myocardial Infarction (n=722, MIGEN)[9], Adult Macular Degeneration (n=486**,** AMD)[10], and Multiple Sclerosis (n=247**,** MS)[11].

b. The CANADA collection is described elsewhere[12]. Cases were recruited from the Toronto area, and based on clinical, serological and radiological data were diagnosed with RA in accordance with 1987 American College of Rheumatology criteria. Subjects diagnosed with RA at an age of 16 years or younger were excluded from the study. Some of the control samples (n=378) were healthy white individuals recruited from the Toronto area who had no history of rheumatoid arthritis or other inflammatory disease. Additional healthy controls (n=1,094) were obtained from a lung cancer study[12,13].

c. The CANADA II collection is described elsewhere[12]. These samples consisted of independent cases and healthy controls recruited recruited from the Toronto and Halifax areas. Cases were identified based on clinical, serological and radiological data in accordance with 1987 American College of Rheumatology criteria. Controls for the replication study were also recruited in Toronto and Halifax who had no history of rheumatoid arthritis or other inflammatory disease.

d. The Epidemiological Investigation of Rheumatoid Arthritis provided a second collection of cases and controls (EIRA-II). This collection was described in detail elsewhere[14]. All cases had RA consistent with 1987 ACR criteria and were CCP positive; they were recruited from Sweden. Healthy controls were similarly recruited from Sweden.

e. The Genomics Collaborative Initiative (GCI) samples are a collection of RA cases recruited from Rheumatology specialty clinics in North America and are described in greater detail elsewhere[15]. All patients were RF positive, and fulfilled 1987 ACR criteria. All cases were matched to healthy controls on the basis of age (within five years), gender, and grandparental country (or region) of origin.

f. The Genetics Network Rheumatology Amsterdam (GENRA) provided a collection of Dutch cases recruited through from the greater Amsterdam region. The cases are described in detail elsewhere[16,17]. Cases were recruited from the outpatient rheumatology clinics of the VU university medical center, the Jan van Breemen institute, and the AMC/University of Amsterdam, all situated in the Amsterdam region in the Netherlands. All cases had RA consistent with 1987 ACR criteria, and only those patients that were CCP positive were used in this study. Healthy controls were recruited from blood donors from the same region.

g. Cases and controls of Dutch origin were recruited at Leiden University Medical Center (LUMC). This collection is described in detail elsewhere[18,19]. All cases were RF or CCP positive and fulfilled 1987 ACR

criteria. Controls were healthy individuals who had been recruited at Leiden University Medical Center as part of a separate study on deep vein thromboses[20].

h. The Nurses Health Study (NHS) collection is described in detail elsewhere[21]. Cases and controls were drawn from the Nurses Health Study and Nurses Health Study II, prospective studies of >200,000 female nurses, that have been followed for as long as 30 years. Potential RA cases were identified with a screening questionnaire, followed by detailed chart review by board certified rheumatologists. All cases fulfilled 1987 ACR criteria. Each case was matched to a control by year of birth, menopausal status, and postmenopausal hormone use. We genotyped all control samples, and all cases with confirmed positive RF or CCP.

i. The North American Rheumatoid Arthritis Consortium provided a collection of samples that we used in replication (NARAC-II). These samples were used as a replication cohort for another separate study and are described in detail elsewhere[12,14]. These were samples drawn from specialty clinics from across North America. All samples were CCP positive and met 1987 ACR criteria. Additional healthy controls were drawn from the New York Cancer Project.

j. The North American Rheumatoid Arthritis Consortium provided a third collection of samples that we used in replication (NARAC-III)[12]. These were samples drawn from specialty clinics from across North America. All samples were CCP positive and met 1987 ACR criteria. Cases in this collection were constituted from (1) singleton cases collected at the

Feinstein Institute, (2) cases collected by Dr. Tuulikki Sokka as part of an early onset RA cohort (ERATER)[22], (3) cases contributed by Dr. Ted Mikuls at the University of Nebraska from the VARA cohort, (4) cases collected by Dr. Lindsey Criswell at the University of California San Francisco, (5) RA cases collected as part of an ongoing cohort study of first degree relatives of RA patients (Michael Holers, PI), and (6) cases that are members of multiplex families with multiple autoimmune diseases in the MADGC collection[23]. We obtained publicly available controls from three different groups (1) shared healthy controls from Study 66 and Study 67 from the Illumina Genotype Control Database (www.illumina.com), and (2) a collection of Parkinson's cases and healthy controls recruited for a Parkinsons Disease study[24].

k. United Kingdom Rheumatoid Arthritis Genetics (UKRAG) collection consists of cases and controls recruited from throughout the United Kingdom and is described in detail elsewhere[25]. Cases and controls were recruited from Manchester, Aberdeen, Leeds, Sheffield, London, and Oxford. We selected a subset of cases that were either postive for RF or CCP. All cases fulfilled the 1987 American College of Rheumatology classification criteria. Healthy controls were recruited from 5 of the same 6 centers (cases only recruited from London).

C. Genotyping and Data Processing

a. **Genotyping.** We genotyped each patient collection for the 22 SNPs selected by GRAIL. We designed a single Sequenom iPlex pool and genotyped EIRA-I, EIRA-II, GENRA samples at the Broad Institute (in Cambridge, MA). We employed the same pool design to genotype

NARAC-II cases and controls at the National Institutes of Arthritis, Musculoskeletal, and Skin Diseases (NIAMS, Bethesda, MD). We designed a separate Sequenom iPlex pool and genotyped the same 22 SNPs in the CANADA-II collection at the Analytic Genetics Technology Centre (Toronto, Canada). We designed a separate Sequenom iPlex pool and genotyped the same 22 SNPs in the UKRAG collection at the Arthritis Research Campaign (arc)–Epidemiology Unit, University of Manchester (Manchester, United Kngdom). We genotyped the 22 SNPs in the GCI and LUMC collections using the kinetic PCR platform[26] at Celera Diagnostics (Alameda, CA). We genotyped these 22 SNPs in the NHS collection using the BioTrove multiplex SNP genotyping assay at the Nurses Health Study (Boston, MA). We obtained genotype data generated at the Broad Institute for these SNPs from previously generated GWA data on the Affymetrix 6.0 platform for BRASS cases and controls; we extracted genotypes or proxies for the 22 SNPs of interest. We genotyped NARAC-III cases with the Illumina 317K array at the Feinstein Institute; we obtained publicly available genotype data on the same platform for shared controls after an official application to a Parkinson's Disease consortium and Illumina Genotype Control Database (www.illumina.com). We genotyped CANADA cases and controls with the Illumina 370K array at Illumina in San Diego, CA. We extracted or imputed genotypes for the 22 SNPs of interest for the NARAC-III and CANADA collections using IMPUTE[2].

b. **SNP Proxies.** Under certain circumstances, where an assay failed or was unavailable, we utilized proxy SNPs in lieu of the selected SNP. All proxy SNPs were in strong LD with the selected SNP with r^2 =1.0 in CEU population of the Phase II Hapmap. We used the following proxies for

selected subsets of patient collections: rs12405671 (for rs11586238); rs12465751 (for rs13393256), rs12569358 (for rs10919563); rs1675766 (for rs2614394); rs2056626 (for rs1773560); rs3821236 (for rs11893432); rs4755453 (for rs540386); rs4839491 (for rs4272626); rs6809087 (for rs4535211); rs7110197 (for rs2276418); rs7257871 (for rs3176767); rs7426056 (for rs1980422); rs7529225 and rs7539468 (for rs12746613); and rs9360720 (for rs9359049).

c. **Quality Control.** For each collection we applied stringent quality control criteria. We required that each SNP pass the following criteria for each collection separately: (1) genotype missing rate < 10%, (2) minor allele frequency > 1%, and (3) Hardy-Weinberg equilibrium with $p>10^{-3}$. We also excluded individuals with data missing for > 10% of SNPs passing quality control.

D. Correcting for case-control stratification. Since the CANADA-II, EIRA-II, GCI, GENRA, LUMC, NHS, and UKRAG collections consisted of cases and controls that were drawn from well-matched populations or were already matched on epidemiological factors, we did not pursue further strategies to correct for potential case-control stratification. The CANADA, BRASS, NARAC-II, NARAC-II collections each included shared controls - we therefore used additional matching based on ancestry informative markers to correct for potential stratification.

a. We matched BRASS cases to AMD, MIGEN, and MS controls using ancestry informative markers from available genome-wide SNP data. We used available data from the Affymetix 6.0 from the Broad Institute that has been used to genotype these samples, and selected a subset of SNPs

passing stringent case control criteria. We used the resulting set of 681,637 SNPs to define genetic eigenvectors with Eigenstrat[27]. Eigenstrat first removed distinct genotypic outliers. We observed significant case-control stratification along the first two eigenvectors. Both correlated with the lactase (LCT) region, known to stratify heavily across different European populations. To match cases and controls we utilized the following strategy: (1) Iterated through cases randomly, (2) For each case we selected the closest control that had not yet been selected in the Euclidean space of the top two Eigenvectors, and (3) Iterated until a total of 3 controls were selected for each case. The resulting collection of cases and controls demonstrated minimal case-control stratification.

b. Similarly for NARAC-II we used data for 760 ancestry informative SNPs for the cases and controls[28]. These markers had been selected for their efficiency in separating northern and southern European populations. We used this data to run Eigenstrat. Eigenstrat removed distinct genotypic outliers. Significant stratification was observed only along the first eigenvector. We used a similar strategy to that described above for the BRASS collection, except we used only one eigenvector and selected only 1.5 controls per case.

c. A similar strategy was also used from the NARAC-III GWA study. Genotype data from GWAS studies were available for 269,771 SNPs that had passed quality control in each of the case and control sub-collections in NARAC-III. These SNPs were used as ancestry informative markers. We used Eigenstrat to remove genetically distinct markers, and then defined genetic principal components. We observed significant case-

control stratification along the top two eigenvectors. We applied the strategy described above to match controls to remaining cases; in this case we selected 1.5 controls per case based on matching on the top two eigenvectors. This analysis of this collection and the resulting samples are identical as those reported elsewhere[1].

d. For the CANADA study, we used a similar strategy. Genotype data was available for GWAS for 269,771 SNPs that passed strict quality control criteria. These SNPs were used as ancestry informative markers. We used Eigenstrat to remove genetically distinct markers, and then defined genetic principal components. We observed significant case-control stratification along the top two eigenvectors. We applied the strategy described above to match controls to remaining cases; in this case we selected 2.5 controls per case based on matching on the top two eigenvectors.

E. Statistical Analysis

a. **Removing Duplicate Samples.** We wanted to be certain that there were duplicate samples (1) within individual studies or (2) across studies that recruited from the same country. To identify duplicates we used genetic data from this study. The 22 SNPs from this study alone, especially after considering failed SNP assays, were generally not specific enough to confidently identify duplicates. Therefore we also used data on an additional 18-53 SNPs passing quality control genotyped for a previous study[1] for six collections (EIRA-II [28 SNPs], GENRA [28 SNPs], GCI [18 SNPs], LUMC [18 SNPs], NARAC-II [52 SNPs], NHS [53 SNPs]). For UKRAG, we used a panel of an additional 118 SNPs passing quality

control genotyped for previous studies[29-36] in addition to the 17 for this study. For CANADA-II we only had available data for the 20 SNPs genotyped for this study available to identify duplicates. For a subset of collections we used available genome-wide data (BRASS, EIRA, NARAC, WTCCC, CANADA, NARAC-III). For these data sets we used high confidence imputations (>0.99 with IMPUTE[2]) where it was necessary to compare to SNPs that were not available on the platform. We grouped collections by country: Canada (CANADA, CANADA-II), United Kingdom (WTCCC, UKRAG), The Netherlands (GENRA, LUMC), Sweden (EIRA, EIRA-II), and the United States (NARAC, NARAC-II, NARAC-III, NHS, BRASS, GCI). We then used the '--genome' option in Plink[37] to identify genetically identical samples within each study and also across studies recruited from the same country. If a pair of identical samples were identified, one of the two samples were removed. In total, we removed 86 cases and 43 controls from the replication samples as a result of this step.

b. **Assessing significance of SNPs in replication.** For each SNP we conducted an 11 strata one-sided CMH[38] statistical test to assess significance of allelic association in replication across all replication samples. We also calculated a CMH odds ratio. Each stratum consisted of an individual collection. A traditional two sided CMH score provides a score, *c*, that is distributed under the random model according to the χ^2

$$z = \sqrt{c} \bullet sign\left[\log\left(OR_{ma}\right) \bullet \log\left(OR_{rep}\right)\right] \qquad (1)$$

distribution. To transform the two-sided CMH replication score, to a one sided replication score, we applied eq. (1) where OR_{ma} is the CMH meta-analysis odds ratio, and OR_{rep} is the CMH replication odds ratio. Under the

null, the *z* value is distributed as a normal distribution with mean zero, and variance one. So if the odds ratios are consistent between the replication and the meta-analysis the *z* value is positive, otherwise the *z* value is negative. We considered those SNPs that obtained a *p*-value threshold of 0.0023 (=0.05/22) as successfully replicated.

c. **Assessing significance of SNPs in joint analysis.** For each SNP we conducted a stratified two-sided CMH statistical test to assess significance of allelic association across all available samples from GWAS meta-analysis and replication collections. In the original publication of the meta-analysis[1] we addressed case-control stratification by subdividing NARAC-I into 396 strata and EIRA-I into 165 strata based on identity-by-state clustering. The WTCCC samples were placed into a single strata. So meta-analysis samples contributed a total of 562 strata. Independent replication samples form this study contributed an additional 11 strata for a total of 573 strata. In joint analysis, we considered a *p*-value threshold of 5 x 10^{-8} overwhelmingly significant, and indicative of an RA associated allele.

d. **Breslow-Day test of heterogeneity.** For each of the SNPs, we calculated a 14 strata Breslow-Day statistic to assess heterogeneity of effect across multiple patient collections[39]. Each patient collection, either from meta-analysis or replication, was placed in a single stratum.

e. **Testing for interactions.** We compiled data for the 22 selected SNPs and also for 17 validated RA SNPs from 16 loci (2 SNPs from the *TNFAIP3* locus, see **Table 1**, main text) for datasets where genome-wide data was

available (NARAC-I, NARAC-III, EIRA-I, EIRA-II, WTCCC, CANADA). For each individual we defined key variables (1) a binary variable indicating case-control status, (2) five binary indicator variables instantiating the patient collection the individual was obtained from, (3) 39 SNP variables ranging from 0-2 indicating the number of minor alleles that the individual has. For each SNP pair we built a logistic regression model to predict case-control status based on the indicator variables and the two SNP variables (a total of 7 variables and an intercept). We tested whether the log-liklihood of the model was significantly improved by adding an additional multiplicative pairwise interaction term for those two SNPs. We conducted a total of 741 (=(39*38)/2) tests. An interaction term was considered significant only if $p<6.7x10^{-5}$ (=0.05/741). For all pairwise tests, we observed that the addition of an interaction test did not significantly improve log-liklihood ($p>0.005$).

f. **Conditional Analysis.** In a single case where the SNP of interest was close to a known validated RA risk locus, we conducted conditional analysis. This situation applied to *CD28* (rs1980422) and *CTLA4* (rs3087243). We used the same formalism as described in the previous section to define a two SNP model across seven patient collections from which genome-wide data was available (WTCCC, CANADA-I, EIRA I, EIRA-II, NARAC-I, NARAC-III, BRASS). We define the model as before with six indicator variables defining patient collection and SNP variables ranging from 0-2 indicating the minor allele count for an individual. We tested whether removing any single SNP from the model significantly worsened the log-liklihood score. We observed that inclusion of both SNPs to the model over a baseline model resulted in a significant improvement

(p=7.6 x 10^{-11} , 'Two Locus' model). Removal of *CTLA4* signficantly worsened the model (p=2.1 x 10^{-7} , 'Single Locus/CD28' model), as did removal of *CD28* (p=3.4 x 10^{-3} , 'Single Locus/CTLA4' model).

References

1. Raychaudhuri, S. et al. Common variants at *CD40* and other loci confer risk of rheumatoid arthritis. *Nat Genet* **40**, 1216-23 (2008).
2. 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**, 906-13 (2007).
3. de Bakker, P.I. et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* **17**, R122-8 (2008).
4. Raychaudhuri, S. et al. Identifying Relationships Among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions. *PLOS Genetics* **5**, e1000534 (2009).
5. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **36**, D13-21 (2008).
6. Arnett, F.C. et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum* **31**, 315-24 (1988).
7. Isenberg, D. *Oxford textbook of rheumatology*, xxi, 1278 p., [63] p. of plates (Oxford University Press, Oxford ; New York, 2004).
8. Sato, M. et al. The validity of a rheumatoid arthritis medical records-based index of severity compared with the DAS28. *Arthritis Res Ther* **8**, R57 (2006).
9. Kathiresan, S. et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* **41**, 334-41 (2009).
10. Neale, B.M. et al. A Genome-Wide Scan of Advanced Age-Related Macular Degeneration Suggests a Novel Role of Lipase-C. **In Review**(2009).
11. De Jager, P.L. et al. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet* **41**, 776-82 (2009).
12. Gregersen, P.K. et al. REL, encoding a member of the NF-κB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nat Genet* **41**, 820-3 (2009).
13. Amos, C.I. et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* **40**, 616-22 (2008).
14. Plenge, R.M. et al. TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. *N Engl J Med* **357**, 1199-209 (2007).

15. 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**, 330-7 (2004).

16. Nielen, M.M. et al. Antibodies to citrullinated human fibrinogen (ACF) have diagnostic and prognostic value in early arthritis. *Ann Rheum Dis* **64**, 1199-204 (2005).

17. Wijbrandts, C.A. et al. The clinical response to infliximab in rheumatoid arthritis is in part dependent on pre-treatment TNF{alpha} expression in the synovium. *Ann Rheum Dis* (2007).

18. 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).

19. Wesoly, J. et al. Association of the PTPN22 C1858T single-nucleotide polymorphism with rheumatoid arthritis phenotypes in an inception cohort. *Arthritis Rheum* **52**, 2948-50 (2005).

20. van der Meer, F.J., Koster, T., Vandenbroucke, J.P., Briet, E. & Rosendaal, F.R. The Leiden Thrombophilia Study (LETS). *Thromb Haemost* **78**, 631-5 (1997).

21. Costenbader, K.H., Chang, S.C., De Vivo, I., Plenge, R. & Karlson, E.W. Genetic polymorphisms in PTPN22, PADI-4, and CTLA-4 and risk for rheumatoid arthritis in two longitudinal cohort studies: evidence of gene-environment interactions with heavy cigarette smoking. *Arthritis Res Ther* **10**, R52 (2008).

22. Sokka, T. & Pincus, T. An Early Rheumatoid Arthritis Treatment Evaluation Registry (ERATER) in the United States. *Clin Exp Rheumatol* **23**, S178-81 (2005).

23. 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**, 561-71 (2005).

24. Fung, H.C. et al. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol* **5**, 911-6 (2006).

25. Thomson, W. et al. Rheumatoid arthritis association at 6q23. *Nat Genet* (2007).

26. Germer, S., Holland, M.J. & Higuchi, R. High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res* **10**, 258-66 (2000).

27. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).

28. Seldin, M.F. et al. European population substructure: clustering of northern and southern populations. *PLoS Genet* **2**, e143 (2006).

29. Plant, D. et al. A re-evaluation of three putative functional single nucleotide polymorphisms in rheumatoid arthritis. *Ann Rheum Dis* **68**, 1373-5 (2009).

30. Orozco, G. et al. Combined effects of three independent SNPs greatly increase the risk estimate for RA at 6q23. *Hum Mol Genet* **18**, 2693-9 (2009).

31. Orozco, G. et al. Association of CD40 with rheumatoid arthritis confirmed in a large UK case-control study. *Ann Rheum Dis* (2009).

32. Burr, M.L. et al. PADI4 genotype is not associated with rheumatoid arthritis in a large UK Caucasian Population. *Ann Rheum Dis* (2009).

33. Barton, A. et al. 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**, 2518-22 (2009).

34. Barton, A. et al. Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat Genet* **40**, 1156-9 (2008).

35. Barton, A. et al. Re-evaluation of putative rheumatoid arthritis susceptibility genes in the post-genome wide association study era and hypothesis of a key pathway underlying susceptibility. *Hum Mol Genet* **17**, 2274-9 (2008).

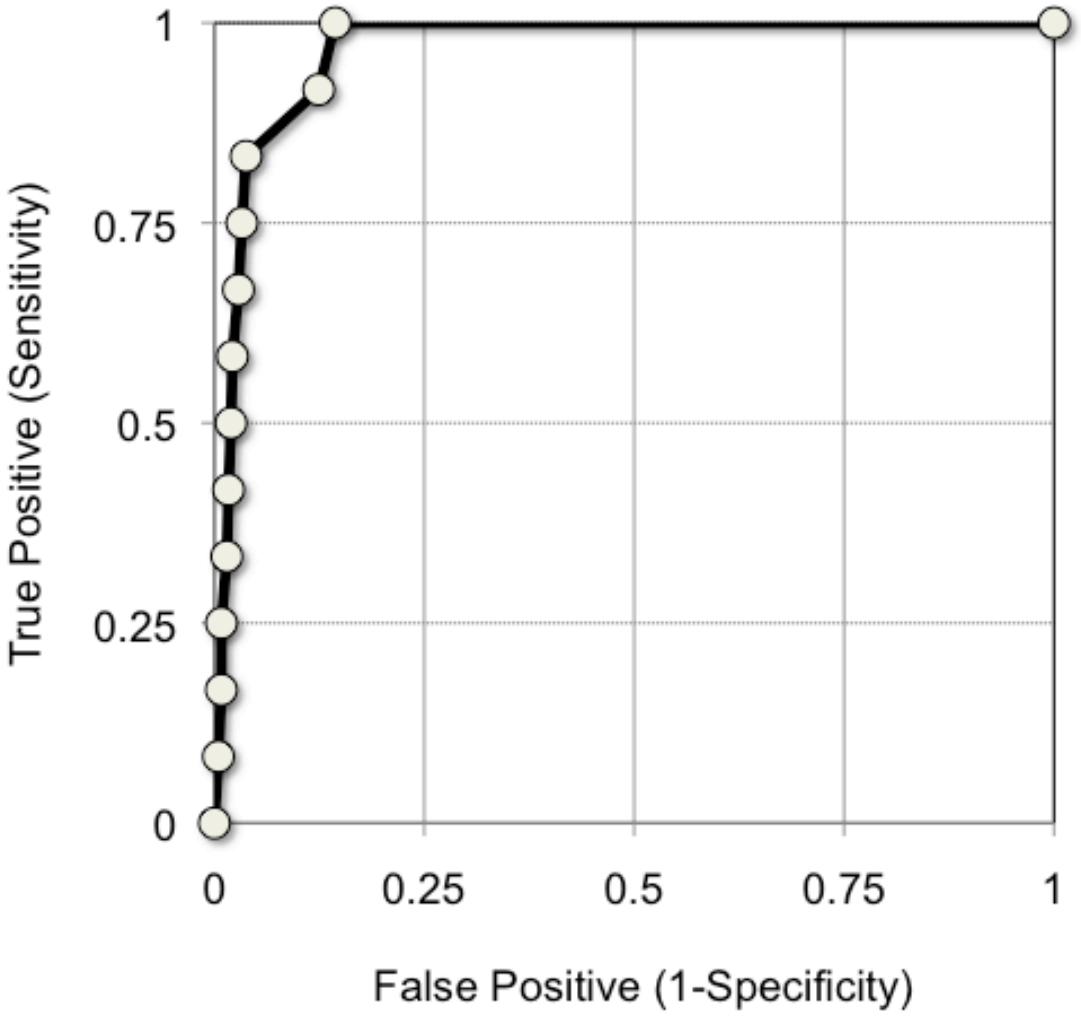
36. Thomson, W. et al. Rheumatoid arthritis association at 6q23. *Nat Genet* **39**, 1431 - 1433 (2007).

37. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

38. Kuritz, S.J., Landis, J.R. & Koch, G.G. A general overview of Mantel-Haenszel methods: applications and recent developments. *Annu Rev Public Health* **9**, 123-60 (1988).

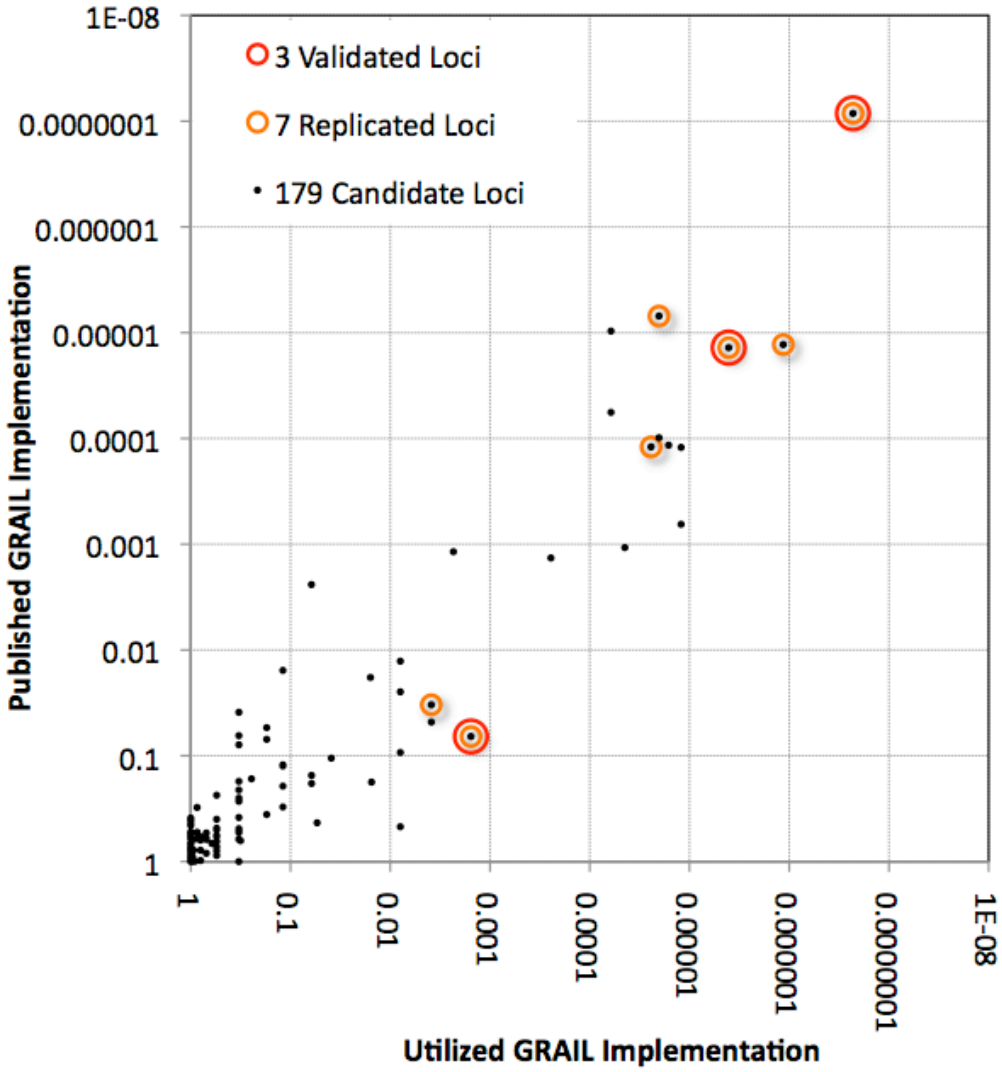
39. Breslow, N.E. Elementary methods of cohort analysis. *Int J Epidemiol* **13**, 112-5 (1984).

Supplementary Figure 1.



Supplementary Figure 1. Receiver-operator curve for GRAIL's ability to identify novel RA loci. We selected 12 loci associated with RA, not included in GRAIL's literature database. We scored them against other independent validated RA SNPs. For comparison we also scored 10,000 random SNPs. For multiple p_{text} thresholds we calculated sensitivity (true positive rate) using these 12 loci and specificity (1 – false positive rate) using 10,000 random SNPs.

Supplementary Figure 2.



Supplementary Figure 2. GRAIL score differences based on implementation. Here we plot the GRAIL scores for the implementation used in this paper (x-axis) and the published implementation (Raychaudhuri et al *PLOS Genetics* 2009) which uses a slightly different statistical model to identify functional connectivity. We note that scores are highly correlated (non-parametric Spearman correlation=0.84). Out of 179 loci, 144 score >0.1 with both implementations.

Supplementary Table 1.

Validated RA Locus	Representative Allele (SNPs)	GRAIL Score	Candidate Gene
1p13.2*	rs2476601	0.0043	*PTPN22*
1p36.13*	rs2240340	0.0012	*PADI4*
1p36.32	rs3890745	1.00E-06	*TNFRSF14*
1q23.3	rs6682654	0.0012	*CD244*
2q33.2*	rs3087243	3.00E-07	*CTLA4*
2q32.3	rs7574865	1.40E-04	*STAT4*
4q27	rs6822844	1.10E-05	*IL2 IL21*
6q23.3	rs10499194, rs6920220	9.10E-05	*TNFAIP3*
6p21.32 (MHC class II)*	rs6457620	0.0043	*HLA-DRA*
7q21.2	rs42041	0.25	*CDK6*
9q33.2	rs3761847	2.30E-04	*TRAF1*
9p13.3	rs2812378	4.50E-05	*CCL21*
10p15.1	rs4750316	0.0018	*PRKCQ*
12q13.3	rs1678542	0.15	*B4GALNT1*
20q13.12	rs4810485	8.40E-06	*CD40*
22q12.3	rs3218253	6.10E-04	*IL2RB*

Supplementary Table 1. Known associated RA risk loci and their GRAIL score.

Here we scored each of the sixteen known associated loci against the other loci with GRAIL. The first two columns describe the locus and a published SNP representing the locus. The GRAIL score is listed in the third column. In the fourth column we list the most connected nearby gene identified by GRAIL. *Loci known prior to December 2006 - these were excluded from the analysis described in the main text, since their association to RA was included in publications contained in the GRAIL text database.

Supplementary Table 2

SNP	Chr	Meta-Analysis *p*	GRAIL *p*	Gene	Previously Genotyped
rs4272626	1	3.5E-04	0.0079	*NHLH2*	
rs11586238	1	2.0E-04	2.0E-05	*CD2, IGSF2, CD58*	
rs12746613	1	9.1E-04	2.4E-05	*FCGR2A*	
rs1773560	1	4.4E-04	6.1E-05	*CD247*	
rs10919563	1	3.8E-04	4.0E-06	*PTPRC*	
rs13393256	2	6.9E-04	6.1E-05	*TTC7A*	
rs7579737	2	8.2E-04	1.6E-05	*IL1RL1*	
rs1980422	2	4.2E-05	2.3E-07	*CD28*	Assay Failed
rs4535211	3	4.4E-04	0.0079	*PLCL2*	
rs231707	4	6.0E-04	1.2E-05	*TNIP2*	
rs9359049	6	2.7E-05	0.0079	*CD109*	
rs548234	6	3.4E-04	0.0016	*PRDM1*	
rs394581	6	5.6E-04	0.0039	*TAGAP*	
rs10282458	7	9.1E-04	0.0039	*RARRES2*	
rs7041422	9	4.7E-04	0.0023	*IFNB1*	
rs540386	11	6.1E-04	1.1E-06	*RAG1, TRAF6*	
rs2276418	11	4.0E-04	2.0E-05	*CD3G*	
rs2614394	12	9.8E-05	1.2E-05	*IRAK4*	Stage 1 only
rs9564915	13	4.3E-04	0.00024	*PIBF1*	
rs7234029	18	1.9E-04	0.0079	*PTPN2*	
rs892188	19	4.6E-05	4.4E-05	*ICAM1, ICAM3*	Assay Failed
rs3176767	19	1.0E-04	4.4E-05	*ICAM1, ICAM3*	

Supplementary Table 2. 22 Candidate SNPs identified by GRAIL. We identified 22 SNPs out of 179 with compelling GRAIL scores. In the first two columns we list information about the SNP. In the third column we list the meta-analysis *p*-value; we

selected only snps with $p < 0.001$. The third column lists the GRAIL score; we list only those SNPs with *GRAIL p* <0.01. In the fifth column we list the gene most connected to other RA associated loci as identified by GRAIL. In the final column we list efforts to genotype these SNPs in the original meta-analysis publication (Raychaudhuri et at 2008).

Supplementary Table 3.

	Case Collection	Control Collection	Case Origin	Case Antibody Status	Cases	Controls	Genotyping Platform	Case-Controls Stratification
Meta-Analysis	Epidemiological Investigation of Rheumatoid Arthritis (EIRA)	EIRA	Sweden	100% CCP+	660	658	Illumina 317K	Epidemiologically Matched, Identity By State Clustering
3,393 Cases; 12,460 Controls	North American Rheumatoid Arthritis Consortium (NARAC)	Shared Controls (New York Cancer Project, New York City)	North America	100% CCP+	873	1196	Illumina 550K	Identity By State Clustering
	Wellcome Trust Case Control Consortium (WTCCC)	Shared Controls (multiple non-autoimmune diseases from WTCCC)	United Kingdom	80% CCP+, 84% RF +	1860	10606	Affymetrix 500K	Geographically Matched
Replication	Brigham Rheumatoid Arthritis Sequential Study (BRASS)	Shared Controls (Unaffected individuals from 3 Studies)	Boston, USA	100% CCP+	485	1455	Affymetrix 6.0	Case-Control Matching with GWAS data
	CANADA	Controls from Toronto, Canada and **Shared Controls** (Unaffected individuals from 1 Study)	Toronto, Canada	100% CCP+	589	1472	Illumina 370K	Case-Control Matching with GWAS data
	CANADA-II	CANADA II	Toronto & Halifax, Canada	100% CCP+	695	1233	Sequenom iPlex	Geographically Matched
7.957 Cases; 11,958 Controls	EIRA-II	EIRA-II	Sweden	100% CCP+	437	407	Sequenom iPlex	Epidemiologically Matched
	Genetics Network Rheumatology Amsterdam (GENRA)	GENRA	Amsterdam, The Netherlands	100% CCP+	549	1219	Sequenom iPlex	Geographically Matched
	Genomics Collaborative Initiative (GCI)	GCI	North America	100% RF+	460	465	Kinetic PCR	Epidemiologically Matched
	Leiden University Medical Center (LUMC)	LUMC	Leiden, The Netherlands	100% RF+ or CCP+	311	541	Kinetic PCR	Geographically Matched
	NARAC-II	Shared Controls (New York Cancer Project, New York City)	North America	100% CCP+	487	733	Sequenom iPlex	Case-Control Matching with Ancestry informative markers
	NARAC-III	Shared Controls (Publically available)	North America	100% CCP+	865	1303	Illumina 317K	Case-Control Matching with GWAS data
	Nurses Health Study (NHS)	NHS	North America	100% RF+ or CCP+	294	506	Biotrove OpenArray	Epidemiologically Matched
	United Kingdom Rheumatoid Arthritis Genetics (UKRAG)	UKRAG	United Kingdom	100% RF+ or CCP+	2785	2624	Sequenom iPlex	Geographically Matched

Supplementary Table 3. Patient collections. The GWAS meta-analysis derived from three collections. The replication set derived from eleven patient 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 list the genotyping technology used to type SNPs of interest. Finally, we specify the strategy used to correct for case-control population stratification.

Supplementary Table 4.

SNP	CHR	Gene	BRASS (*z*)	CANADA (*z*)	CANADA II (*z*)	EIRA II (*z*)	GENRA (*z*)	GCI (*z*)	LUMC (*z*)	NARAC II (*z*)	NARAC III (*z*)	NHS (*z*)	UKRAG (*z*)	All (*z*)	All (*p*)
rs4272626	1p13.1	*NHLH2*	-1.04	1.72	1.78	0.62	0.13	1.99	-2.03	N/A	-0.06	1.04	0.94	1.73	0.042
rs11586238	1p13.1	*CD2, IGSF2, CD58*	0.98	2.51	2.86	0.31	1.36	-0.03	0.76	3.54	2.09	-0.35	1.22	4.69	1.4E-06
rs12746613	1q23.3	*FCGR2A*	0.38	1.05	N/A	-1.38	-1.34	-0.21	0.43	1.78	2.41	0.62	2.86	2.85	0.0022
rs1773560	1q24.2	*CD247*	-0.85	1.60	0.27	-0.72	1.11	0.31	1.47	1.04	1.27	0.36	0.71	2.02	0.021
rs10919563	1q31.3	*PTPRC*	-0.48	1.66	1.15	1.77	1.06	1.73	0.30	1.23	1.82	-0.07	1.34	3.47	0.00026
rs13393256	2p21	*TTC7A*	0.27	0.36	-1.55	-0.45	2.31	-0.50	-0.04	-0.81	0.57	N/A	N/A	0.142	0.44
rs7579737	2q12.1	*IL1RL1*	-1.62	-0.48	-0.93	-1.51	-0.86	-0.66	0.16	-0.29	-1.29	1.07	0.65	-1.51	0.93
rs1980422	2q33.2	*CD28*	0.52	0.87	1.93	0.81	1.24	1.67	1.69	-0.99	2.09	0.36	3.02	4.43	4.7E-06
rs4535211	3p24.3	*PLCL2*	1.58	1.48	0.64	-0.98	1.04	0.34	1.61	0.55	-0.12	0.37	0.77	2.18	0.015
rs231707	4p16.3	*TNIP2*	1.80	0.83	0.63	0.07	0.79	0.69	1.22	1.88	0.09	-0.63	-0.56	1.66	0.048
rs9359049	6q13	*CD109*	-0.31	-1.43	-1.47	-0.47	0.32	N/A	N/A	N/A	-0.63	-1.12	0.05	-1.53	0.94
rs548234	6q21	*PRDM1*	1.71	1.10	1.07	1.26	1.30	2.99	-0.72	1.49	-0.19	1.60	2.35	4.22	1.2E-05
rs394581	6q25.3	*TAGAP*	-0.21	0.70	0.44	1.21	1.28	0.29	2.92	1.22	1.27	-0.80	2.73	3.62	0.00015
rs10282458	7q36.1	*RARRES2*	1.15	0.33	-1.36	-1.35	0.85	-1.04	1.14	2.89	-0.24	N/A	N/A	0.75	0.23
rs7041422	9p21.3	*IFNB1*	0.44	-0.07	0.78	0.82	0.08	0.07	0.86	-0.00	-0.33	-0.13	N/A	0.71	0.24
rs540386	11p12	*RAG1, TRAF6*	1.27	-0.38	0.19	-0.19	2.54	0.74	0.46	0.70	3.37	2.17	0.28	3.14	0.00083
rs2276418	11q23.3	*CD3G*	0.64	-0.54	-1.59	-0.28	0.96	-0.95	0.43	-0.43	0.43	2.19	2.54	1.43	0.077
rs2614394	12q12	*IRAK4*	-0.14	-1.65	N/A	N/A	-0.60	-0.68	-1.24	N/A	0.34	N/A	N/A	-1.52	0.94
rs9564915	13q22.1	*PIBF1*	-0.12	0.07	-1.39	0.45	-0.01	-1.83	-0.31	-0.14	0.43	0.52	2.25	0.62	0.27
rs7234029	18p11.21	*PTPN2*	0.65	0.84	0.74	1.91	1.26	1.00	0.31	-1.33	-0.40	1.09	1.39	2.24	0.013
rs892188	19p13.2	*ICAM1, ICAM3*	0.34	1.18	0.90	0.74	0.88	-1.82	1.21	0.19	0.41	1.44	N/A	1.73	0.041
rs3176767	19p13.2	*ICAM1, ICAM3*	-0.35	1.00	-0.30	1.61	-0.05	0.70	0.93	-0.16	-0.08	1.20	0.65	1.31	0.095

Supplementary Table 4. Genotyping results by individual patient collection. For each of the 22 SNPs that we pursued in this study, we list information describing the SNP in the first three columns including the SNP ID, location, and candidate genes identified by GRAIL. For each patient collection we list a z-score suggesting the strength of association for that SNP. A positive z-score suggests that the direction of association is consistent with that of the original meta-analysis. Boxes are highlighted in yellow if z>1.63, corresponding to $p<0.05$. Boxes are grey if genotype data was not available (N/A) due to failure to pass quality control criteria. In the final two columns we list the aggregate z-score from all replication samples and the corresponding one-tailed *p*-value - for 13 out of 22 SNPs z>1.63.

Supplementary Table 5.

Locus	Result	Baseline Model	Single Locus/CD28	Single Locus/CTLA4	Two Locus
rs3087243 / CTLA4	Odds Ratio			0.875 (0.837 - 0.916)	0.886 (0.846 - 0.927)
rs1980422 / *CD28*	Odds Ratio		1.108 (1.053 - 1.166)		1.081 (1.026 - 1.138)
	-2 x LL	24966.8	24951.5	24933.1	24924.5
	Model vs Two Locus (*p*)	7.6×10^{-11}	2.1×10^{-7}	3.4×10^{-3}	
	Model vs Baseline (*p*)		8.8×10^{-5}	6.2×10^{-9}	7.6×10^{-11}

Supplementary Table 5. Conditional Analysis between CTLA4 and CD28 loci. We conducted conditional analysis on the *CTLA4* locus and the *CD28* locus. We defined four logistic regression models: (1) a baseline model with only information about the patient collection, (2) a single locus model with a CD28 allele count, (3) a single locus model with a CTLA4 allele count, and (4) a two locus model with allele counts for both SNPs. For each model with list the Odds Ratios ($=e^{\beta}$) and their 95% confidence intervals for each parameter. We also list the -2 x log likelihood (LL) fit of their data. The significance in model change can be calculated from the difference in -2 x LL between two models - under the null this difference is distributed according to the chi-square distribution. For each model, we have calculated the significance of the model's improvement in likelihood compared to that of the Baseline Model (Model vs Baseline), and also the significance of the Two Locus model's improvement in likelihood over each model (Model vs Two Locus). Given a model with one locus already, addition of the second locus improves likelihood of the data significantly ($p=2.1 \times 10^{-7}$ for addition of *CTLA4* and $p=3.4 \times 10^{-3}$ for addition of *CD28*).