

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

Soumya Raychaudhuri^{1-3*}, Brian P Thomson², Elaine F Remmers⁴, Stephen Eyre⁵, Anne Hinks⁵, Candace Guiducci², Joseph J Catanese⁶, Gang Xie⁷, Eli A Stahl¹, Robert Chen¹, Lars Alfredsson⁸, Christopher I Amos⁹, Kristin G Ardlie², BIRAC Consortium²⁸, Anne Barton⁵, John Bowes⁵, Noel P Burt², Monica Chang⁶, Jonathan Coblyn¹, Karen H Costenbader¹, Lindsey A Criswell¹⁰, J Bart A Crusius¹¹, Jing Cui¹, Phillip L De Jager^{2,12}, Bo Ding⁸, Paul Emery¹³, Edward Flynn⁵, Pille Harrison¹⁴, Lynne J Hocking¹⁵, Tom W J Huizinga¹⁶, Daniel L Kastner⁴, Xiayi Ke⁵, Fina A S Kurreeman^{1,16}, Annette T Lee¹⁷, Xiangdong Liu⁷, Yonghong Li⁶, Paul Martin⁵, Ann W Morgan¹³, Leonid Padyukov¹⁸, David M Reid¹⁵, Mark Seielstad¹⁹, Michael F Seldin²⁰, Nancy A Shadick¹, Sophia Steer²¹, Paul P Tak²², Wendy Thomson⁵, Annette H M van der Helm-van Mil¹⁶, Irene E van der Horst-Bruinsma²³, Michael E Weinblatt¹, Anthony G Wilson²⁴, Gert Jan Wolbink^{25,26}, Paul Wordsworth¹⁴, YEAR Consortium²⁸, David Altschuler^{2,3}, Elizabeth W Karlson¹, Rene E M Toes¹⁶, Niek de Vries²², Ann B Begovich^{6,27}, Katherine A Siminovitch⁷, Jane Worthington⁵, Lars Klareskog¹⁸, Peter K Gregersen¹⁷, Mark J Daly^{2,3} & Robert M Plenge^{1,2}

To discover new rheumatoid arthritis (RA) risk loci, we systematically examined 370 SNPs from 179 independent loci with $P < 0.001$ in a published meta-analysis of RA genome-wide association studies (GWAS) of 3,393 cases and 12,462 controls¹. We used Gene Relationships Across Implicated Loci (GRAIL)², a computational method that applies statistical text mining to PubMed abstracts, to score these 179 loci for functional relationships to genes in 16 established RA disease loci^{1,3-11}. We identified 22 loci with a significant degree of functional connectivity. We genotyped 22 representative SNPs in an independent set of 7,957 cases and 11,958 matched controls. Three were convincingly validated: *CD2-CD58* (rs11586238, $P = 1 \times 10^{-6}$ replication, $P = 1 \times 10^{-9}$ overall), *CD28* (rs1980422, $P = 5 \times 10^{-6}$ replication, $P = 1 \times 10^{-9}$ overall) and *PRDM1* (rs548234, $P = 1 \times 10^{-5}$ replication, $P = 2 \times 10^{-8}$ overall). An additional four were replicated ($P < 0.0023$): *TAGAP* (rs394581, $P = 0.0002$ replication, $P = 4 \times 10^{-7}$ overall), *PTPRC* (rs10919563, $P = 0.0003$ replication, $P = 7 \times 10^{-7}$ overall), *TRAF6-RAG1* (rs540386, $P = 0.0008$ replication, $P = 4 \times 10^{-6}$ overall) and *FCGR2A* (rs12746613, $P = 0.0022$ replication, $P = 2 \times 10^{-5}$ overall). Many of these loci are also associated to other immunologic diseases.

Rheumatoid arthritis (RA) is a chronic autoimmune disease characterized by inflammatory polyarthritis¹². Genetic studies have now identified multiple risk alleles for autoantibody-positive RA within the *MHC* region, a *PTPN22* missense allele, and risk alleles in 14

other loci (Table 1)^{1,3-11}. Most RA risk loci contain multiple genes, and currently the causal genes within most risk loci are unknown. However, most RA risk loci contain at least one plausible biological candidate gene involved in immune regulation, and these genes suggest an important set of processes involved in RA pathogenesis. For example, risk alleles highlight genes involved in T-cell activation by antigen-presenting cells (class II *MHC* region, *PTPN22*, *STAT4* and *CTLA4*), the NF- κ B signaling pathway (*CD40*, *TRAF1*, *TNFSF14* and *TNFAIP3*, and the recent report of *REL*¹³), citrullination (*PADI4*), natural killer cells (*CD244*) and chemotaxis (*CCL21*).

Based on these observations, we hypothesized that as-yet-undiscovered autoantibody-positive RA risk loci might also contain genes with functions similar to those of genes in known risk loci. Therefore, known RA risk loci can be used to prioritize SNPs for replication from GWAS in independent samples (Fig. 1), especially those SNPs with modest statistical support.

To objectively quantify the degree of functional similarity between genes within candidate loci identified from GWAS and genes within validated RA risk loci, we used a published functional genomics method, GRAIL (Gene Relationships Across Implicated Loci)². GRAIL quantifies functional similarity between genes by applying established statistical text mining methods¹⁴ to a database of 250,000 published scientific abstracts about human and model-organism genes. For each candidate locus, GRAIL identifies the gene with the greatest number of observed relationships to other genes. GRAIL estimates the statistical significance of the number of observed relationships with a null model in which relationships between genes near SNPs occur by

*A full list of author affiliations appears at the end of the paper.

Table 1 Sixteen validated RA loci used in functional analyses

Validated RA locus	Representative allele (SNPs)	Genes within associated regions
1p13.2 ^a	rs2476601	<i>PTPN22</i> , <i>AP4B1</i> , <i>RSBN1</i> , <i>BCL2L15</i> , <i>DCLRE1B</i> , <i>MAGI3</i> , <i>PHTF1</i>
1p36.13 ^a	rs2240340	<i>PADI3</i> , <i>PADI4</i>
1p36.32	rs3890745	<i>PANK4</i> , <i>MMEL1</i> , <i>PLCH2</i> , <i>HES5</i> , <i>TNFRSF14</i>
1q23.3	rs6682654	<i>LY9</i> , <i>CD244</i>
2q33.2 ^a	rs3087243	<i>ICOS</i> , <i>CTLA4</i>
2q32.3	rs7574865	<i>STAT1</i> , <i>GLS</i> , <i>STAT4</i>
4q27	rs6822844	<i>IL2</i> , <i>IL21</i> , <i>ADAD1</i> , <i>KIAA1109</i>
6q23.3	rs10499194, rs6920220	<i>OLIG3</i> , <i>TNFAIP3</i>
6p21.32 ^a (MHC class II)	rs6457620, DRB1*0401, *0101	<i>HLA-DRA</i> , <i>HLA-DQB1</i> , <i>BTNL2</i> , <i>HLA-DQA1</i> , <i>HLA-DRB5</i> , <i>HLA-DRB1</i>
7q21.2	rs42041	<i>PEX1</i> , <i>FAM133B</i> , <i>GATAD1</i> , <i>CDK6</i>
9q33.2	rs3761847	<i>PHF19</i> , <i>CEP110</i> , <i>TRAF1</i> , <i>RAB14</i> , <i>C5</i>
9p13.3	rs2812378	<i>CCL21</i>
10p15.1	rs4750316	<i>RBM17</i> , <i>PFKFB3</i> , <i>PRKCK</i>
12q13.3	rs1678542	<i>DTX3</i> , <i>METTL1</i> , <i>AVIL</i> , <i>DDIT3</i> , <i>XRC-C6BP1</i> , <i>MBD6</i> , <i>GLI1</i> , <i>CYP27B1</i> , <i>KIF5A</i> , <i>GEFT</i> , <i>CTDSP2</i> , <i>MARS</i> , <i>CDK4</i> , <i>AGAP</i> , <i>DCTN2</i> , <i>TSPAN31</i> , <i>FAM119B</i> , <i>MARCH9</i> , <i>TSFM</i> , <i>B4GALNT1</i> , <i>OS9</i> , <i>PIP4K2C</i> , <i>ARHGAP9</i> , <i>SLC26A10</i>
20q13.12	rs4810485	<i>SLC12A5</i> , <i>NCOA5</i> , <i>CD40</i>
22q12.3	rs3218253	<i>IL2RB</i>

The 16 established RA loci (column 1), a representative SNP from each (column 2) and all of the genes in LD with the SNP (column 3). For each SNP, the gene in boldface is the one that GRAIL selected as the most functionally connected gene when that locus was scored against the 15 other validated risk loci.

^aLoci discovered before December 2006.

random chance. This significance score, P_{text} , represents the output GRAIL score. GRAIL is already able to effectively identify functional interconnectivity between genes within the previously known RA loci (Fig. 2); it might also be able to establish connections between these 16 loci and as-yet-undiscovered RA risk loci.

Because GRAIL might show variable performance across different phenotypes, we wanted to carefully quantify its predictive ability in RA before using it to prioritize SNPs for replication. To estimate GRAIL's ability to distinguish true RA loci from spurious associations, we examined 12 RA risk loci discovered since 2006 (Table 1,

Supplementary Table 1). The current GRAIL implementation is based on PubMed abstracts published before December 2006. As these 12 risk loci were discovered since this date, they constitute a representative set to evaluate GRAIL's performance. In a 'leave-one-out' analysis, we used GRAIL to score each of these loci for functional relationships to the other 15 validated RA risk loci. A total of 10 of the 12 loci obtained GRAIL scores of $P_{\text{text}} < 0.01$. This analysis suggests that at this P_{text} threshold, GRAIL has an ~83% true positive rate (or sensitivity). To assess the false positive rate of this same P_{text} threshold, we modeled spurious loci by sampling 10,000 random SNPs from the Affymetrix 500K array; we scored these SNPs against all 16 RA loci. Of the randomly selected SNPs, 5.4% scored $P_{\text{text}} < 0.01$; this corresponds to a specificity of ~95%. Assessment of true and false positive rates at different cutoffs revealed an area under the curve (or C statistic) of 0.97 (Supplementary Fig. 1). We note that if a large number of candidate SNPs are screened in a study, this might still result in a large number of false positives.

Next, we attempted to identify new RA risk loci from a set of SNPs with modest evidence of association from our recent GWAS meta-analysis of 3,392 affected individuals (cases) and 12,462 controls¹. In our original study, we genotyped SNPs with $P < 10^{-4}$ in the meta-analysis and found that 6 out of 31 SNPs replicated in our independent samples. However, many RA risk alleles have modest effects (for example, OR < 1.2) and will be missed at that significance threshold. We therefore expected that some SNPs at $P < 0.001$ may be risk alleles. After excluding SNPs that were known, validated RA risk loci, we identified a total of 370 SNPs from 179 independent regions that obtained $P < 0.001$ (Online Methods and Supplementary Note). The total number of SNPs observed at this threshold was consistent with the approximate number of SNPs expected by chance, suggesting that the majority of these SNPs represent spurious associations and should not be reproducible in an independent case-control study.

For each of the 179 candidate loci, we selected the single SNP with the strongest evidence from the GWAS meta-analysis and then scored it against the 16 validated RA risk loci with GRAIL. If all 179 SNPs were spurious, then ~10 should score $P_{\text{text}} < 0.01$ based on the estimated 5.4% false positive rate. However, 22 of the 179 (12.3%) scored $P_{\text{text}} < 0.01$ (Fig. 3a, Supplementary Table 2). This represented a significant enrichment compared to random sets of 179 SNPs ($P = 3.3 \times 10^{-4}$ by simulation). We therefore expected that of this select subset of 22 SNPs, as many as half might represent true RA risk alleles.

To identify which of these 22 SNPs represented true RA risk loci, we genotyped them in an independent validation study of 7,957 cases and 11,958 controls from 11 collections from Europe and North America (Supplementary Table 3). All cases met the 1987 American College of

Figure 1 Using Gene Relationships Across Implicated Loci (GRAIL) to prioritize candidate RA SNPs. We selected a set of candidate SNPs to pursue in an independent genotyping experiment by starting with all SNPs that obtained $P < 0.001$ in an independent GWAS meta-analysis. Then, for each candidate SNP, GRAIL identified the genomic region in LD and identified overlapping genes. It then checked to see how many other loci already known to be associated with disease contained functionally related genes. SNPs representing those candidate loci with significantly related genes were forwarded for genotyping in large numbers of independent case-control samples.

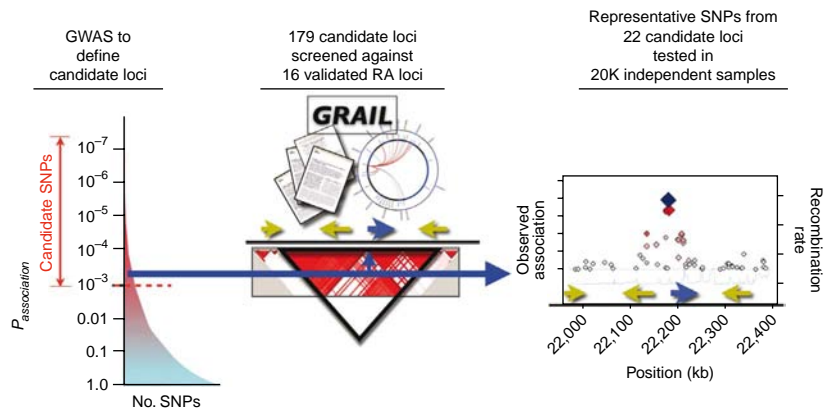
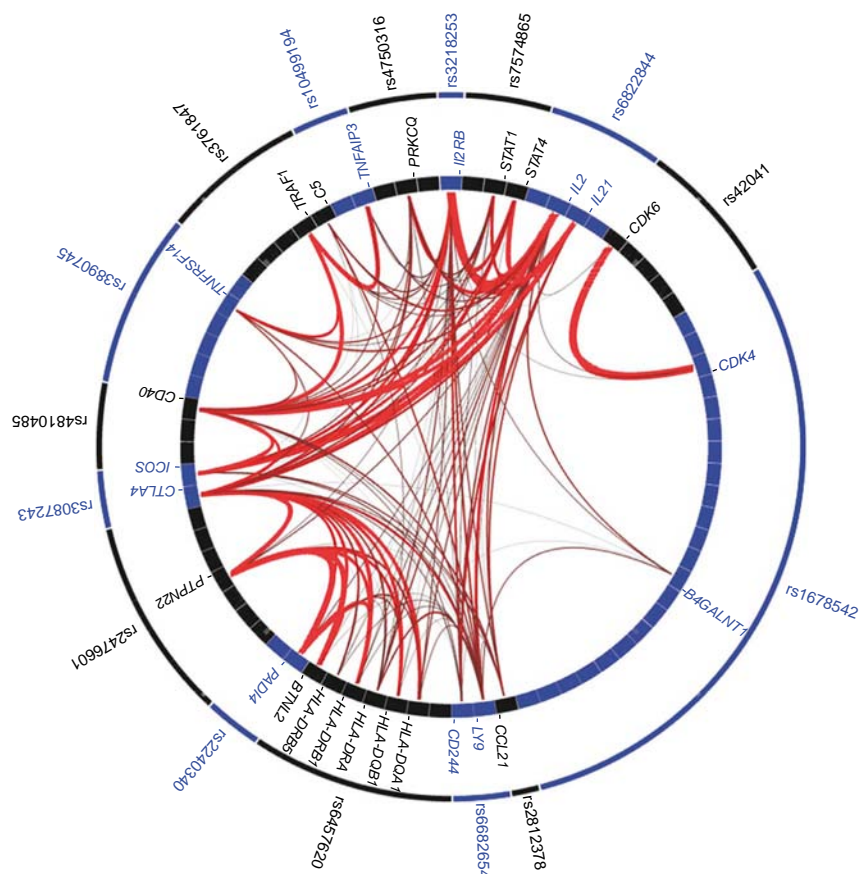


Figure 2 GRAIL identified interconnectivity among genes in RA loci. The known RA-associated SNPs are along the outer ring; the internal ring represents the genes near each SNP (as listed in **Table 1**); each box in the internal ring represents an individual gene. We illustrate the literature-based functional connectivity between these genes with lines drawn between them—the redder and thicker the lines are, the stronger the connectivity between the genes is. RA SNPs implicate a small number of highly connected genes—those genes are indicated by labeled boxes.



Rheumatology classification criteria¹⁵ or were diagnosed by a board-certified rheumatologist and were seropositive for disease-specific autoantibodies (anticyclic citrullinated peptide (CCP) antibody or rheumatoid factor (RF)). All individuals were self-described as “white” and of European ancestry. We assessed association with a Cochran-Mantel-Haenszel (CMH) stratified association statistic¹⁶. For each SNP, we calculated a z score, where a $z > 0$ indicates the same allele confers risk in both the replication and the meta-analysis samples. To interpret statistical significance, we used a Bonferroni-corrected one-tailed P value of 0.0023 (calculated as $0.05/22$, $z > 2.83$). Additionally, we calculated the overall association P value across all samples (GWAS meta-analysis plus replication).

Notably, of the 22 SNPs examined, 19 (86%) obtained $z > 0$ (Fig. 3b). If these SNPs represented spurious associations, then only about half should have $z > 0$; the probability of such a positive skew

in the number of SNPs with $z > 0$ by chance alone is $P_{\text{skew}} = 0.0005$, suggesting a high likelihood of a large number of true RA risk loci within this set of 22 SNPs.

Of the 22 SNPs selected by GRAIL, 13 obtained nominal levels of association to RA at $P < 0.05$ (corresponding to $z > 1.65$), whereas no more than 2 might be expected by chance alone. More compellingly, seven SNPs achieved a Bonferroni-corrected level of significance in replication ($P < 0.0023$, $z > 2.83$).

When we aggregated both GWAS meta-analysis and replication genotype data (Table 2, Supplementary Table 4), we observed the strongest evidence of association to RA at rs11586238 on 1p13.1 near the *CD2* and *CD58* genes ($P = 1.4 \times 10^{-6}$ replication, $P = 1.0 \times 10^{-9}$ overall), at rs1980422 on 2q33.2 near *CD28* ($P = 4.7 \times 10^{-6}$ replication, $P = 1.3 \times 10^{-9}$ overall) and at rs548234 on 6q21 near *PRDM1* ($P = 1.2 \times 10^{-5}$ replication, $P = 2.1 \times 10^{-8}$ overall). Based

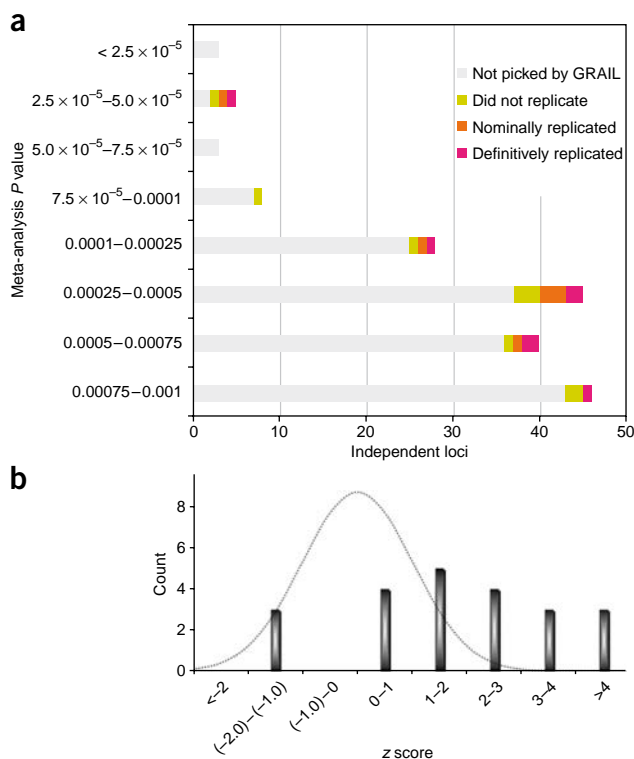


Figure 3 GRAIL identifies new RA risk loci that replicate when genotyped in independent case-control samples. (a) GRAIL identified 22 SNPs among the 179 candidate SNPs with $P < 0.001$ in a GWAS meta-analysis. This is a histogram of the 179 SNPs as a function of their GWAS meta-analysis P value. Gray bars represent the 157 SNPs that were not selected and colored bars represent the 22 SNPs that were selected; purple indicates SNPs that replicated convincingly in follow-up genotyping ($P < 0.0023$), orange indicates nominally associated SNPs in follow-up genotyping ($P < 0.05$), and yellow indicates genotyped SNPs without any independent evidence of association. (b) Enrichment of SNPs with z scores > 2 in replication samples. For each of the 22 SNPs tested, we calculated a one-sided CMH z -score statistic from our two-staged replication data. A z score of 0 corresponds to a $P = 0.5$, a z score of 1.65 to $P = 0.05$, and a z score of 2.83 to $P = 0.0023$. For a random collection of unassociated SNPs, the histogram should approximate a normal distribution (dotted line).

Table 2 SNPs tested for RA susceptibility

ID	SNP			Meta-analysis						Replication			Joint			
	Chr	Pos (HG18)	Gene(s)	Allele		P	OR	Minor allele		P	OR	Minor allele		P	OR	Breslow-Day
				Major	Minor			Control	Case			Control	Case			
Replicated loci (uncorrected $P < 0.0023$)																
rs11586238 ^a	1p13.1	117,064,661	<i>CD2</i> , <i>IGSF2</i> , <i>CD58</i>	C	G	2.0×10^{-4}	1.14	0.237	0.260	1.4×10^{-6}	1.12	0.228	0.254	1.0×10^{-9}	1.13	0.29
rs1980422 ^a	2q33.2	204,318,641	<i>CD28</i>	T	C	4.2×10^{-5}	1.16	0.230	0.255	4.7×10^{-6}	1.11	0.237	0.255	1.3×10^{-9}	1.13	0.81
rs548234 ^a	6q21	106,674,727	<i>PRDM1</i>	T	C	3.4×10^{-4}	1.12	0.328	0.351	1.2×10^{-5}	1.10	0.323	0.343	2.1×10^{-8}	1.11	0.66
rs394581 ^a	6q25.3	159,402,509	<i>TAGAP</i>	T	C	5.6×10^{-4}	0.89	0.302	0.269	1.5×10^{-4}	0.92	0.286	0.270	3.8×10^{-7}	0.91	0.63
rs10919563 ^a	1q31.3	196,967,065	<i>PTPRC</i>	G	A	3.8×10^{-4}	0.84	0.128	0.108	2.6×10^{-4}	0.90	0.132	0.117	6.7×10^{-7}	0.88	0.64
rs540386	11p12	36,481,869	<i>RAG1</i> , <i>TRAF6</i>	C	T	6.1×10^{-4}	0.86	0.142	0.130	8.3×10^{-4}	0.91	0.145	0.130	3.9×10^{-6}	0.89	0.08
rs12746613 ^a	1q23.3	159,733,666	<i>FCGR2A</i>	C	T	9.1×10^{-4}	1.16	0.120	0.133	0.0022	1.10	0.124	0.130	1.5×10^{-5}	1.12	0.25
Nominally associated loci (uncorrected $P < 0.05$)																
rs7234029 ^a	18p11.21	12,867,060	<i>PTPN2</i>	A	G	1.9×10^{-4}	1.16	0.158	0.179	0.013	1.06	0.164	0.172	4.4×10^{-5}	1.10	0.61
rs4535211	3p24.3	17,048,001	<i>PLCL2</i>	G	A	4.4×10^{-4}	0.90	0.489	0.457	0.015	0.96	0.474	0.461	8.9×10^{-5}	0.94	0.524
rs1773560	1q24.2	165,688,387	<i>CD247</i>	A	G	4.4×10^{-4}	0.90	0.421	0.385	0.021	0.96	0.414	0.401	1.5×10^{-4}	0.94	0.74
rs892188	19p13.2	10,270,793	<i>ICAM1</i> , <i>ICAM3</i>	C	T	4.6×10^{-5}	1.13	0.378	0.409	0.041	1.05	0.393	0.401	4.3×10^{-5}	1.08	0.21
rs4272626	1p13.1	116,149,950	<i>NHLH2</i>	C	T	3.5×10^{-4}	1.12	0.359	0.388	0.042	1.04	0.354	0.362	1.9×10^{-4}	1.07	0.07
rs231707	4p16.3	2,664,183	<i>TNIP2</i>	G	A	6.0×10^{-4}	1.14	0.178	0.195	0.048	1.05	0.172	0.184	5.3×10^{-4}	1.08	0.23
Loci that failed to replicate																
rs2276418	11q23.3	117,735,474	<i>CD3G</i>	A	T	4.0×10^{-4}	1.16	0.142	0.161	0.077	1.04	0.155	0.155	9.5×10^{-4}	1.08	0.11
rs3176767	19p13.2	10,310,751	<i>ICAM1</i> , <i>ICAM3</i>	T	G	1.0×10^{-4}	1.15	0.224	0.245	0.09	1.03	0.229	0.233	6.9×10^{-4}	1.07	0.60
rs10282458	7q36.1	149,676,235	<i>RARRES2</i>	G	A	9.1×10^{-4}	1.12	0.259	0.282	0.23	1.02	0.260	0.266	4.4×10^{-3}	1.06	0.045
rs7041422	9p21.3	21,034,021	<i>IFNB1</i>	T	G	4.7×10^{-4}	1.12	0.300	0.331	0.24	1.02	0.297	0.301	4.4×10^{-3}	1.06	0.86
rs9564915	13q22.1	72,223,143	<i>PIBF1</i>	A	G	4.3×10^{-4}	1.12	0.319	0.341	0.27	1.01	0.317	0.315	0.008	1.05	0.14
rs13393256	2p21	47,140,263	<i>TTC7A</i>	C	A	6.9×10^{-4}	1.13	0.210	0.227	0.44	1.00	0.221	0.221	0.014	1.06	0.14
rs7579737	2q12.1	102,353,793	<i>IL1RL1</i>	A	G	8.2×10^{-4}	0.89	0.307	0.274	0.93	1.04	0.295	0.308	0.483	0.99	0.023
rs2614394	12q12	42,568,433	<i>IRAK4</i>	G	A	9.8×10^{-5}	0.81	0.098	0.082	0.94	1.08	0.099	0.105	0.06	0.94	0.002
rs9359049	6q13	74,758,649	<i>CD109</i>	T	A	2.7×10^{-5}	1.27	0.068	0.081	0.94	0.94	0.079	0.071	0.14	1.05	0.0155

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 (OR) with respect to the minor allele. The next four columns list similar results for replication genotyping; significance is reported based on stratified one-tailed CMH statistic. The next three columns summarize joint (overall) analysis results; significance is reported based on stratified two-tailed CMH statistic across all 14 patient collections (3 from the meta-analysis and 11 from the replication study). The final column lists the Breslow-Day Test for heterogeneity of odds ratios across all 14 collections.

^aThese SNPs are close to other loci already associated to autoimmune disease.

on conservative estimates of genome-wide significance ($P = 5 \times 10^{-8}$), these SNPs represent confirmed RA risk alleles.

Four additional loci replicated; however, in aggregate analysis of GWAS meta-analysis and replication genotype data, no loci exceeded a conservative estimate of significance. We observed evidence of association at rs394581 on 6q25.3 near *TAGAP* ($P = 1.5 \times 10^{-4}$ replication, $P = 3.8 \times 10^{-7}$ overall), rs10919563 on 1q31.3 within a *PTPRC* intron ($P = 2.6 \times 10^{-4}$ replication, $P = 6.7 \times 10^{-7}$ overall), rs540386 on 11p12 within a *TRAF6* intron ($P = 8.3 \times 10^{-4}$ replication, $P = 3.9 \times 10^{-6}$ overall) and rs12746613 on 1q23.3 near *FCGR2A* ($P = 2.2 \times 10^{-3}$ replication, $P = 1.5 \times 10^{-5}$ overall). These SNP associations likely represent true RA loci, but additional genotyping will be necessary for definitive confirmation.

Notably, many of the SNPs picked by GRAIL that validated in independent genotyping were not those with strongest evidence of association in the initial GWAS meta-analysis (Fig. 3a). That is, prioritization based purely on meta-analysis P values might have missed many of these associations. For example, rs12746613 (*FCGR2A*) was ranked 163 of 179 and rs540386 (*RAG1-TRAF6*) was ranked 110. Of the five SNPs that we genotyped with the most significant GRAIL P_{text}

scores, three replicated and one showed nominal evidence of association; only rs2614394 (*IRAK4*) showed no evidence of association.

Many of these seven alleles further implicate genomic regions already associated with autoimmune diseases (Table 3). At this point, none of these RA risk alleles correspond perfectly to any previously established autoimmune allele, but in some cases, fine mapping of the region in multiple diseases could clarify the relationships between the alleles. The rs12746613 SNP in *FCGR2A* is 13 kb away from a missense SNP in *FCGR2A* that has been associated with systemic lupus erythematosus^{17,18}; these two SNPs are in the same LD block ($r^2 = 0.19$, $D' = 1.0$). The rs394581 SNP is located in the 5' untranslated region of *TAGAP* and is 17 kb away from a SNP associated with celiac disease and with type 1 diabetes^{19,20}; these two SNPs are in partial LD ($r^2 = 0.32$, $D' = 0.73$). The rs10919563 SNP in *PTPRC* is 35 kb away from a rare (~1% allele frequency) nonsynonymous SNP that alters splicing of *PTPRC*²¹; there have been inconsistent reports that the latter allele is associated with multiple sclerosis²²⁻²⁴. We also note that the rs7234029, a *PTPN2* intronic SNP, is 41 kb away from a SNP associated with both type 1 diabetes and celiac disease²⁰; these two alleles are in the same LD block ($r^2 = 0.14$, $D' = 1.0$). The rs548234 SNP is

Table 3 SNPs near other alleles associated with autoimmune diseases

SNP			Published SNP			Proximity		
ID	Chr	Gene	ID	Gene	Disease associations	Distance (kb)	r^2	D'
rs12746613	1q23.3	<i>FCGR2A</i>	rs1801274	<i>FCGR2A</i>	Systemic lupus erythematosus	12.7	0.19	1.00
rs394581	6q25.3	<i>TAGAP</i>	rs1738074	<i>TAGAP</i>	Celiac disease, type 1 diabetes	16.5	0.32	0.73
rs10919563	1q31.3	<i>PTPRC</i>	rs17612648	<i>PTPRC</i>	Multiple sclerosis	34.5	—	—
rs7234029	18p11.21	<i>PTPN2</i>	rs478582	<i>PTPN2</i>	Type 1 diabetes	41.1	0.14	1.00
rs1980422	2q33.2	<i>CD28</i>	rs3087243	<i>CTLA4</i>	Type 1 diabetes, RA	128.5	0.04	0.40
rs548234	6q21	<i>PRDM1</i>	rs7746082	<i>PRDM1</i>	Crohn's disease	132.8	0.01	0.08
rs11586238	1p13.1	<i>CD2</i>	rs2300747	<i>CD58</i>	Multiple sclerosis	158.9	0.01	0.29

Seven of the 22 SNPs tested are near loci already associated with autoimmune diseases. The first three columns list the SNPs, cytogenetic location and the likely candidate gene. The next three columns list the published SNP, the attributed gene and the disease associations. The final three columns list the physical distance and measures of LD. For *PTPRC*, the published SNP is rare and LD cannot be accurately assessed.

located 10 kb downstream from the *PRDM1* transcript and is 133 kb away from a SNP previously associated with Crohn's disease²⁵. The rs11586238 SNP is 50 kb upstream of the *CD2* start site but is also near multiple other key immunological genes including *CD58* and *IGSF2*. This SNP is also 159 kb away from a multiple sclerosis-associated SNP within a *CD58* intron^{26,27}.

The rs1980422 SNP is located about 10 kb away from the 3' untranslated region of *CD28* and is 129 kb away from a known RA and type 1 diabetes risk allele in the *CTLA4* region (rs3087243)¹¹. There is minimal LD between these two SNPs ($r^2 = 0.04$, $D' = 0.40$); conditional analysis confirmed that these two SNPs independently confer RA risk (see **Supplementary Table 5**).

These SNP associations continue to clarify critical biological processes involved in RA pathogenesis, including T-cell activation, NF- κ B signaling and B-cell activation and differentiation. The CD2 protein is a co-stimulatory molecule on the surface of natural killer cells and T-cells; signaling through CD2 is mediated by its binding of PTPRC directly²⁸. SNP association to CD28 contributes additional evidence of the role of T-cell activation in disease pathogenesis. TRAF6 is involved in downstream NF- κ B activation; it binds CD40 directly and is a key component of B-cell activation²⁹. Our study has also implicated new processes represented by PRDM1 (also known as BLIMP-1), a transcription factor that regulates terminal differentiation of B-cells into immunoglobulin secreting plasma cells³⁰. Functional studies and resequencing will be required to confirm that these genes are indeed the truly causal genes for RA pathogenesis.

We examined all seven replicated RA SNPs along with known RA risk alleles for epistatic interactions (**Supplementary Note**). Despite the functional relationships between these genes, we found no evidence of significant interactions.

Population stratification could result in spurious associations. However, we were careful for each collection to use either epidemiologically matched samples or ancestry-informative markers to match cases and controls. We further note that our 7 replicated SNP associations showed consistent effects across all 14 collections without evidence of heterogeneity ($P > 0.05$ by Breslow-Day test of heterogeneity, **Table 2**).

In this study, we demonstrated the utility of functional information to prioritize SNPs for replication. We did not predefine pathways but rather used GRAIL to look for genes that had relationships to other validated RA genes. We note that GRAIL is limited in its ability to identify disease genes in entirely new pathways (that is, pathways not suggested by the 16 previously known RA risk loci). Arguably, it is those disease genes that could point to truly new pathogenic mechanisms. Additionally, successful application of GRAIL is contingent on the scientific literature's comprehensive description of relevant gene

relationships. The general application of GRAIL to other diseases will depend critically on the completeness of the validated loci list and the documentation about relevant processes in the literature. Despite these limitations, our study has identified at least three previously unknown RA risk loci and has showed strong evidence for additional risk loci.

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

S.R. is supported by a US National Institutes of Health (NIH) Career Development Award (1K08AR055688-01A1) and an American College of Rheumatology Bridge Grant. 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, the American College of Rheumatology 'Within Our Reach' campaign and holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund. M.J.D. is supported by NIH grants through the U01 (HG004171, DK62432) and R01 (DK083756-1, DK64869) mechanisms. The Broad Institute Center for Genotyping and Analysis is supported by grant U54 RR020278 from the National Center for Research Resources. The Brigham and Women's Hospital Rheumatoid Arthritis Sequential Study (BRASS) Registry is supported by a grant from Millennium Pharmaceuticals and Biogen-Idec. The North American Rheumatoid Arthritis Consortium (NARAC) is supported by the NIH (NO1-AR-2-2263 and RO1 AR44422). This research was also supported in part by the Intramural Research Program of the National Institute of Arthritis, Musculoskeletal and Skin Diseases of the NIH. This research was also supported in part by grants to KAS from the Canadian Institutes for Health Research (MOP79321 and IIN - 84042) and the Ontario Research Fund (RE01061) and by a Canada Research Chair. We acknowledge the help of C.E. van der Schoot for healthy control samples for the Genetics Network Rheumatology Amsterdam (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 patient samples and data. We thank the Myocardial Infarction Genetics Consortium (MIGen) study for the use of genotype data from their healthy controls in our study. The MIGen study was funded by the US NIH and National Heart, Lung, and Blood Institute's SNP Typing for Association with Multiple Phenotypes from Existing Epidemiologic Data (STAMPEED) genomics research program R01HL087676 and a grant from the National Center for Research Resources. We thank the Johanna Seddon Progression of AMD Study, AMD Registry Study, Family Study of AMD, The US Twin Study of AMD and the Age-Related Eye Disease Study (AREDS) for use of genotype data from their healthy controls in our study. We thank D. Hafler and the Multiple Sclerosis collaborative for use of genotype data from their healthy controls recruited at Brigham and Women's Hospital.

AUTHOR CONTRIBUTIONS

S.R., M.J.D., D.A. and R.M.P. designed the study, conducted the statistical analysis, interpreted the primary data and wrote the initial manuscript. All authors contributed to the final manuscript. B.P.T., E.F.R., S.E., A.H., C.G., J.J.C., G.X.,

E.A.S., R.C., N.P.B. and M.S. were involved directly in genotyping samples or extracting genotypes for this study. The BRASS genetic study was coordinated by E.A.S., P.L.d.J., J.C., S.R. and R.M.P. under the direction of M.E.W. and N.A.S. The CANADA genetic study was coordinated by C.I.A., X.L. and G.X. under the direction of K.A.S. The Epidemiological Investigation of Rheumatoid Arthritis (EIRA) genetic study was coordinated by L.A., B.D., L.P. and M.S. under the direction of L.K. The Genomics Collaborative Initiative (GCI) genetic study was coordinated by K.G.A., J.J.C., M.C. and Y.L. under the direction of A.B.B. The GENRA genetic study was coordinated by J.B.A.C., P.P.T., I.E.v.d.H.-B. and G.J.W. under the direction of N.d.V. The Leiden University Medical Center (LUMC) genetic study was coordinated by T.W.J.H., F.A.S.K., Y.L. and A.H.M.v.d.H.-v.M. under the direction of R.E.M.T. The NARAC genetic study was coordinated by E.F.R., C.I.A., M.C., L.A.C., D.L.K., A.T.L. and M.F.S. under the direction of P.K.G. The NHS genetic study was coordinated by K.H.C. and J.C. under the direction of E.W.K. The UK Rheumatoid Arthritis Genetics (UKRAG) genetic study was coordinated by S.E., B.I.R.A.C., A.B., J.B., P.E., E.F., P.H., A.H., L.J.H., X.K., P.M., A.W.M., D.M.R., S.S., W.T., A.G.W., P.W. and Y.E.A.R. under the direction of J.W.

Published online at <http://www.nature.com/naturegenetics/>.

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>.

1. Raychaudhuri, S. *et al.* Common variants at *CD40* and other loci confer risk of rheumatoid arthritis. *Nat. Genet.* **40**, 1216–1223 (2008).
2. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
3. Gregersen, P.K., Silver, J. & Winchester, R.J. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum.* **30**, 1205–1213 (1987).
4. 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).
5. Plenge, R.M. *et al.* Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat. Genet.* **39**, 1477–1482 (2007).
6. Thomson, W. *et al.* Rheumatoid arthritis association at 6q23. *Nat. Genet.* **39**, 1431–1433 (2007).
7. 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).
8. Suzuki, A. *et al.* Functional SNPs in *CD244* increase the risk of rheumatoid arthritis in a Japanese population. *Nat. Genet.* **40**, 1224–1229 (2008).
9. Barton, A. *et al.* Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat. Genet.* **40**, 1156–1159 (2008).

10. 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).
11. 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).
12. Isenberg, D. *Oxford Textbook of Rheumatology* (Oxford University Press, Oxford, UK, and New York, 2004).
13. Gregersen, P.K. *et al.* REL, encoding a member of the NF- κ B family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nat. Genet.* **41**, 820–823 (2009).
14. Raychaudhuri, S. *Computational Text Analysis for Functional Genomics and Bioinformatics* (Oxford University Press, Oxford, 2006).
15. Arnett, F.C. *et al.* The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum.* **31**, 315–324 (1988).
16. 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**, 123–160 (1988).
17. Duits, A.J. *et al.* Skewed distribution of IgG Fc receptor IIa (CD32) polymorphism is associated with renal disease in systemic lupus erythematosus patients. *Arthritis Rheum.* **38**, 1832–1836 (1995).
18. 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).
19. Hunt, K.A. *et al.* Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* **40**, 395–402 (2008).
20. Smyth, D.J. *et al.* Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N. Engl. J. Med.* **359**, 2767–2777 (2008).
21. Tchilian, E.Z. *et al.* The exon A (C77G) mutation is a common cause of abnormal *CD45* splicing in humans. *J. Immunol.* **166**, 6144–6148 (2001).
22. Barcellos, L.F. *et al.* *PTPRC* (CD45) is not associated with the development of multiple sclerosis in U.S. patients. *Nat. Genet.* **29**, 23–24 (2001).
23. Jacobsen, M. *et al.* A point mutation in *PTPRC* is associated with the development of multiple sclerosis. *Nat. Genet.* **26**, 495–499 (2000).
24. Vorechovsky, I. *et al.* Does 77C→G in *PTPRC* modify autoimmune disorders linked to the major histocompatibility locus? *Nat. Genet.* **29**, 22–23 (2001).
25. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
26. De Jager, P.L. *et al.* The role of the *CD58* locus in multiple sclerosis. *Proc. Natl. Acad. Sci. USA* **106**, 5264–5269 (2009).
27. Rubio, J.P. *et al.* Replication of *KIAA0350*, *IL2RA*, *RPL5* and *CD58* as multiple sclerosis susceptibility genes in Australians. *Genes Immun.* **9**, 624–630 (2008).
28. Schraven, B., Samstag, Y., Altevogt, P. & Meuer, S.C. Association of *CD2* and *CD45* on human T lymphocytes. *Nature* **345**, 71–74 (1990).
29. Ishida, T. *et al.* Identification of *TRAF6*, a novel tumor necrosis factor receptor-associated factor protein that mediates signaling from an amino-terminal domain of the *CD40* cytoplasmic region. *J. Biol. Chem.* **271**, 28745–28748 (1996).
30. Calame, K. Activation-dependent induction of *Blimp-1*. *Curr. Opin. Immunol.* **20**, 259–264 (2008).

¹Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, USA. ²Broad Institute, Cambridge, Massachusetts, USA. ³Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. ⁴Genetics and Genomics Branch, National Institute of Arthritis and Musculoskeletal and Skin Diseases, US National Institutes of Health, Bethesda, Maryland, USA. ⁵Arthritis Research Campaign (arc)-Epidemiology Unit, The University of Manchester, Manchester, United Kingdom. ⁶Celera, Alameda, California, USA. ⁷Department of Medicine, University of Toronto, Mount Sinai Hospital and University Health Network, Toronto, Ontario, Canada. ⁸Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. ⁹University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. ¹⁰Rosalind Russell Medical Research Center for Arthritis, Department of Medicine, University of California, San Francisco, California, USA. ¹¹Laboratory of Immunogenetics, Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands. ¹²Department of Neurology, Center for Neurologic Diseases, Brigham and Women's Hospital, Boston, Massachusetts, USA. ¹³National Institute for Health Research-Leeds Musculoskeletal Biomedical Research Unit, Leeds Institute of Molecular Medicine, University of Leeds, United Kingdom. ¹⁴University of Oxford Institute of Musculoskeletal Sciences, Botnar Research Centre, Oxford, United Kingdom. ¹⁵Musculoskeletal and Genetics Section, Division of Applied Medicine, University of Aberdeen, United Kingdom. ¹⁶Department of Rheumatology, Leiden University Medical Centre, Leiden, The Netherlands. ¹⁷The Feinstein Institute for Medical Research, North Shore-Long Island Jewish Health System, Manhasset, New York, USA. ¹⁸Rheumatology Unit, Department of Medicine, Karolinska Institutet at Karolinska University Hospital Solna, Stockholm, Sweden. ¹⁹Genome Institute of Singapore, Singapore. ²⁰Rowe Program in Genetics, University of California at Davis, Davis, California, USA. ²¹Clinical and Academic Rheumatology, Kings College Hospital National Health Service Foundation Trust, Denmark Hill, London, United Kingdom. ²²Clinical Immunology and Rheumatology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. ²³Department of Rheumatology, VU University Medical Center, Amsterdam, The Netherlands. ²⁴School of Medicine and Biomedical Sciences, Sheffield University, Sheffield, United Kingdom. ²⁵Jan van Breemen Institute, Amsterdam, The Netherlands. ²⁶Sanquin Research Landsteiner Laboratory, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. ²⁷Roche Diagnostics, Pleasanton, California, USA. ²⁸A full list of members is provided in the **Supplementary Note**. Correspondence should be addressed to S.R. (soumya@broad.mit.edu) or R.M.P. (rplenge@partners.org).

ONLINE METHODS

Evaluating GRAIL for its ability to identify RA loci. GRAIL is a method that leverages statistical text-mining principles to assess whether putative disease loci harbor genes with functional relationships to genes in other associated disease loci². Two genes are considered similar if the words used to describe them in PubMed abstracts suggest similar functionality. The implementation of GRAIL used here leverages a text database of 250,000 abstracts published before December 2006.

To test the ability of GRAIL to distinguish RA risk loci from spurious associations, we defined a set of true positive loci that were discovered since December 2006; these loci would not be described in the GRAIL text database. We also approximated a set of spurious associations by randomly selecting 10,000 SNPs from the Affymetrix 500K genotyping array. We tested both SNP sets for relationships to known RA-associated loci with GRAIL. Validated SNPs were tested against the other 15 independent loci; spurious SNPs were tested against all 16 loci. The sensitivity was defined as the proportion of true positive associations that GRAIL assigned a $P_{\text{text}} < 0.01$ significance score; the specificity was defined as the proportion of spurious associations that GRAIL assigned a $P_{\text{text}} > 0.01$ significance score.

Selecting nominally associated SNPs for follow-up. To identify SNPs for follow-up, we examined the results of a recently published meta-analysis of three GWAS studies (**Supplementary Table 3**)¹. We examined 336,721 SNPs outside the *MHC* region that passed strict quality control criteria. We identified those SNPs that were nominally associated with RA ($P < 0.001$). We grouped SNPs into independent loci; two SNPs were placed 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 (**Table 1**). We also removed loci with $P < 10^{-4}$ that were genotyped in most available patient collections and had failed to validate in a previous study¹. From the remaining set of independent loci, we selected the single SNP that showed the greatest evidence of association in the published meta-analysis.

Testing SNPs with GRAIL. We tested 179 candidate SNPs using GRAIL for relationships to genes within the 16 independent loci known to be associated with RA. SNPs that obtained compelling GRAIL scores ($P_{\text{text}} < 0.01$) were selected for follow-up investigation. To assess the degree of enrichment among high scoring SNPs, we sampled 100,000 random sets of 179 SNPs and tested these SNP sets with GRAIL. We calculated the proportion of sets with as many or more GRAIL hits to calculate a permutation-based P value. We note that the version of GRAIL that we used is a previous implementation that differs slightly from the published implementation²—results are not substantially affected when those for the same experiment done with the current version of GRAIL (**Supplementary Fig. 2**).

Subject collections. The collections including the study participants with RA and matched controls are described in detail in **Supplementary Table 3** and in the **Supplementary Note**. Each collection consisted only of individuals that were self-described as being “white” and of European descent, and all cases either met the 1987 American College of Rheumatology classification criteria or were diagnosed with RA by board-certified rheumatologists. Informed consent was obtained from each participant, and the Institutional Review Board at each collecting site approved the study.

All cases were autoantibody positive (CCP and/or RF). For most of the collections, matched control samples were collected along with case samples as part of the same study. For some of the collections, where control samples were unavailable, we matched these case collections to shared controls. We used a total of 11 separate case-control collections for replication genotyping: (i) CCP-positive cases from the Brigham Rheumatoid Arthritis Sequential Study (BRASS)³¹ and controls from three separate studies on multiple sclerosis³², age-related macular degeneration (B.M. Neale, J. Fagerness, R. Reynolds, L. Sobrin, M. Parker, S. Raychaudhuri *et al.* unpublished results) and myocardial infarction³³; (ii) CCP-positive cases from the Toronto area (CANADA)¹³ and controls recruited from the same site along with additional controls taken from a disease study of lung cancer³⁴; (iii) CCP-positive cases and controls from Halifax and Toronto (CANADA-II)¹³; (iv) CCP-positive cases from Sweden and epidemiologically matched controls (EIRA-II)³⁵; (v) CCP-positive

Dutch cases and controls collected from the greater Amsterdam region (GENRA)^{36,37}; (vi) North American RF-positive cases and controls matched on gender, age and grandparental country of origin from the Genomics Collaborative Initiative (GCI)⁴; (vii) CCP- or RF-positive Dutch cases and controls from Leiden University Medical Center (LUMC)^{38,39}; (viii) CCP-positive cases drawn from North American clinics and controls from the New York Cancer Project (together this collection is called NARAC-II)^{13,35}; (ix) CCP-positive cases drawn from North American clinics (NARAC-III)¹³ and publicly available controls taken from a Parkinson’s study⁴⁰ and study 66 and 67 of the Illumina Genotype Control Database; (x) 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 use⁴¹; and (xi) CCP- or RF-positive cases recruited at multiple sites in the United Kingdom by the United Kingdom Rheumatoid Arthritis Genetics (UKRAG) collaboration⁶. We used available SNP data from this and previous studies to identify genetically identical samples from the same country; we assumed these represented duplicated individuals and removed them.

Genotyping. A detailed description of the genotyping done is provided in the **Supplementary Note**. All GWAS meta-analysis genotyping was previously described. We genotyped replication samples at the Broad Institute (Cambridge, Massachusetts, USA) using a single Sequenom iPLEX Pool (for the EIRA-II and GENRA collections) and Affymetrix 6.0 (BRASS), the US National Institutes of Health using a single Sequenom iPLEX Pool (NARAC-II), the Analytic Genetics Technology Centre in Toronto using a single Sequenom iPLEX Pool (CANADA-II), the Epidemiology Unit at the University of Manchester using a single Sequenom iPLEX Pool (UKRAG), Celera (Alameda, California, USA) using kinetic PCR⁴² (GCI and LUMC), at the Nurses Health Study in Boston using the BioTrove multiplex SNP genotyping assay (NHS), at the Feinstein Institute (Manhasset, New York) using the Illumina 317K array (NARAC-III); and at Illumina (San Diego, California) using the Illumina 370K array (CANADA). For NARAC-III we additionally obtained publicly available shared controls genotyped on a similar platform from two separate studies. In the cases where whole genome data were available, we either extracted data for the 22 SNPs (BRASS) or used imputation to estimate genotypes for them (CANADA and NARAC-III).

For each collection, we applied stringent quality control criteria. 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 then excluded individuals with data missing for $> 10\%$ of SNPs passing quality control.

Population stratification. For each replication collection, we corrected for possible population stratification by either using only epidemiologically matched samples when cases and controls were drawn from the same population, or matching at least one control for each case based on ancestry-informative markers (see **Supplementary Note** for details). Because the cases in the NHS, GCI, LUMC, EIRA-II, CANADA-II, UKRAG and GENRA collections were well matched to controls, we did not pursue further strategies to correct for population stratification. For the BRASS, NARAC-II, CANADA and NARAC-III collections, we matched cases and controls with ancestry-informative markers and placed them each into a single stratum. For the BRASS cases and shared controls, GWAS data on Affymetrix 6.0 (unpublished data) was available; we used 681,637 SNPs passing strict quality control as ancestry-informative markers. For NARAC-II cases and NYCP shared controls, cases and controls were matched using genotype data on 760 ancestry-informative markers. For the NARAC-III cases and shared controls, we used available Illumina 317K GWAS data for 269,771 SNPs passing stringent quality control criteria. For the CANADA cases and controls, we used available Illumina 317K GWAS 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 (sigma threshold, 6 with 5 iterations) with the software program EIGENSTRAT⁴³. We eliminated vectors that correlated with known structural variants on chromosomes 8 and 17, showed 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 as described elsewhere¹.

Analysis of genetic data. For each SNP, we conducted three statistical tests. First, we conducted a one-sided CMH statistical test across 11 strata to assess whether RA association was reproducible in the replication collections in the same direction as the GWAS meta-analysis. We set our significance threshold, after correcting for 22 hypothesis tests, to be $P < 0.0023$ (calculated by $0.05/22$). Second, we conducted a 573-strata joint analysis across all meta-analysis strata and substrata and replication strata; the 11 replication collections were each placed into their own strata and the meta-analysis samples were partitioned into 562 strata to be consistent with the approach taken in the original analysis to correct for stratification^{1,35}. Third, we calculated a Breslow-Day test of heterogeneity of odds ratios. We performed all analyses in MATLAB (MathWorks).

URLs. Gene Relationships Across Implicated Loci, <http://www.broad.mit.edu/mpg/grail/>; Illumina Genotype Control Database, <http://www.illumina.com>; EIGENSTRAT, <http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm>.

31. 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).
32. 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**, 776–782 (2009).
33. Kathiresan, S. *et al.* Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet.* **41**, 334–341 (2009).
34. 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**, 616–622 (2008).
35. 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).
36. Nielsen, M.M. *et al.* Antibodies to citrullinated human fibrinogen (ACF) have diagnostic and prognostic value in early arthritis. *Ann. Rheum. Dis.* **64**, 1199–1204 (2005).
37. Wijbrandts, C.A. *et al.* The clinical response to infliximab in rheumatoid arthritis is in part dependent on pre-treatment TNF α expression in the synovium. *Ann. Rheum. Dis.* (2007).
38. 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).
39. Wesoly, J. *et al.* Association of the PTPN22 C1858T single-nucleotide polymorphism with rheumatoid arthritis phenotypes in an inception cohort. *Arthritis Rheum.* **52**, 2948–2950 (2005).
40. 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**, 911–916 (2006).
41. 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).
42. Germer, S., Holland, M.J. & Higuchi, R. High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res.* **10**, 258–266 (2000).
43. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

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

Soumya Raychaudhuri^{1-3 ††}, Brian P. Thomson², Elaine F. Remmers⁴, Stephen Eyre⁵, Anne Hinks⁵, Candace Guiducci², Joseph J. Catanese⁶, Gang Xie⁷, Eli A. Stahl¹, Robert Chen¹, Lars Alfredsson⁸, Christopher I. Amos⁹, Kristin G. Ardlie², BIRAC Consortium[†], Anne Barton⁵, John Bowes⁵, Noel P. Burt², Monica Chang⁶, Jonathan Coblyn¹, Karen H. Costenbader¹, Lindsey A. Criswell¹⁰, J. Bart A. Crusius¹¹, Jing Cui¹, Phillip L. De Jager^{2,12}, Bo Ding⁸, Paul Emery¹³, Edward Flynn⁵, Pille Harrison¹⁴, Lynne J. Hocking¹⁵, Tom W. J. Huizinga¹⁶, Daniel L. Kastner⁴, Xiayi Ke⁵, Fina A. S. Kurreeman^{1,16}, Annette T. Lee¹⁷, Xiangdong Liu⁷, Yonghong Li⁶, Paul Martin⁵, Ann W. Morgan¹³, Leonid Padyukov¹⁸, David M. Reid¹⁵, Mark Seielstad¹⁹, Michael F. Seldin²⁰, Nancy A. Shadick¹, Sophia Steer²¹, Paul P. Tak²², Wendy Thomson⁵, Annette H. M. van der Helm-van Mil¹⁶, Irene E. van der Horst-Bruinsma²³, Michael E. Weinblatt¹, Anthony G. Wilson²⁴, Gert Jan Wolbink^{25,26}, Paul Wordsworth¹⁴, YEAR Consortium[†], David Altshuler^{2,3}, Elizabeth W. Karlson¹, Rene E. M. Toes¹⁶, Niek de Vries²², Ann B. Begovich^{6,27}, Katherine A. Siminovitch⁷, Jane Worthington⁵, Lars Klareskog¹⁸, Peter K. Gregersen¹⁷, 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 Emery¹, Professor Philip Conaghan¹, Professor Ann Morgan¹, Dr. Mark Quinn², Dr. Anne-Maree Keenan¹, Dr. Elizabeth Hensor¹

Consultants: Dr. Andrew Gough³, Dr. Michael Green^{2,3}, Dr. Richard Reece⁴, Dr. Lesley Hordon⁵, Dr. Philip Helliwell^{1,6}, Dr. Richard Melsom⁶, Dr. Sheelagh Doherty⁷, Dr. Ade Adebajo⁸, Dr. Andrew Harvey⁹, Dr. Steve Jarrett⁹, Dr. Gareth Huson¹, Dr. Amanda Isdale², Dr. Mike Martin¹, Dr. Zunaid Karim⁹, Professor Dennis McGonagle¹⁰, Dr. Colin Pease¹, Dr. Sally Cox¹

SpRs: Dr. Victoria Bejarano¹, Dr. Jackie Nam¹

Nurses: Claire Brown¹, Christine Thomas¹, David Pickles¹, Alison Hammond¹, Beverley Neville³, Alan Fairclough⁴, Caroline Nunns⁴, Anne Gill², Julie Green², Belinda Rhys-Evans¹, Barbara Padwell¹, Julie Madden¹⁰, Lynda Taylor¹⁰, Sally Smith¹, Heather King¹, Jill Firth⁶, Jayne Heard⁷, Linda Sigsworth⁶

Lab Staff: Diane Corscadden¹, Karen Henshaw¹, Lubna-Haroon Rashid¹, Stephen G Martin¹, James I Robinson¹

¹Section of Musculoskeletal Disease, LIMM, Leeds, UK

²York District Hospital, York, UK

³Harrogate District Hospital, Harrogate, UK

⁴Huddersfield Royal Infirmary, Huddersfield, UK

⁵Dewsbury District and General Hospital, Dewsbury, UK

⁶St Luke's Hospital, Bradford, UK

⁷Hull Royal Infirmary, Hull, UK

⁸Barnsley District General Hospital, Barnsley, UK

⁹Pinderfields General Hospital, Wakefield, UK

¹⁰Calderdale 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 Alexandra 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 Hospital - 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 additional genotyping.

a. Meta-analysis of three GWAS. We had previously conducted a large scale meta-analysis of three GWAS, described in detail elsewhere¹. 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 original 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 algorithm⁴ we tested these 179 SNPs against 16 independent known associated RA risk loci. The implementation 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 organisms⁵. 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 $p_{\text{text}} < 0.01$. 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 ACR⁶ 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 elsewhere⁸. 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, MIGN)⁹, Adult Macular Degeneration (n=486, AMD)¹⁰, and Multiple Sclerosis (n=247, MS)¹¹.
- b.** The CANADA collection is described elsewhere¹². 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 elsewhere¹². These samples consisted of independent cases and healthy controls 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 elsewhere¹⁴. 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 elsewhere¹⁵. 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 thromboses²⁰.

- h.** The Nurses Health Study (NHS) collection is described in detail elsewhere²¹. 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 collection²³. 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 study²⁴.

k. United Kingdom Rheumatoid Arthritis Genetics (UKRAG) collection consists of cases and controls recruited from throughout the United Kingdom and is described in detail elsewhere²⁵. Cases and controls were recruited from Manchester, Aberdeen, Leeds, Sheffield, London, and Oxford. We selected a subset of cases that were either positive 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 Kingdom). We genotyped the 22 SNPs in the GCI and LUMC collections using the kinetic PCR platform²⁶ 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 IMPUTE².

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 Eigenstrat²⁷. 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 controls²⁸. 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 elsewhere¹.

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 study¹ 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 studies²⁹⁻³⁶ 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 IMPUTE²) 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 Plink³⁷ 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 CMH³⁸ 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

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

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-analysis¹ 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 from this study contributed an additional 11 strata for a total of 573 strata. In joint analysis, we considered a p -value threshold of 5×10^{-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 collections³⁹. 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-likelihood 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 $p < 6.7 \times 10^{-5}$ $(=0.05/741)$. For all pairwise tests, we observed that the addition of an interaction test did not significantly improve log-likelihood ($p > 0.005$).

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-likelihood score. We observed that inclusion of both SNPs to the model over a baseline model resulted in a significant improvement

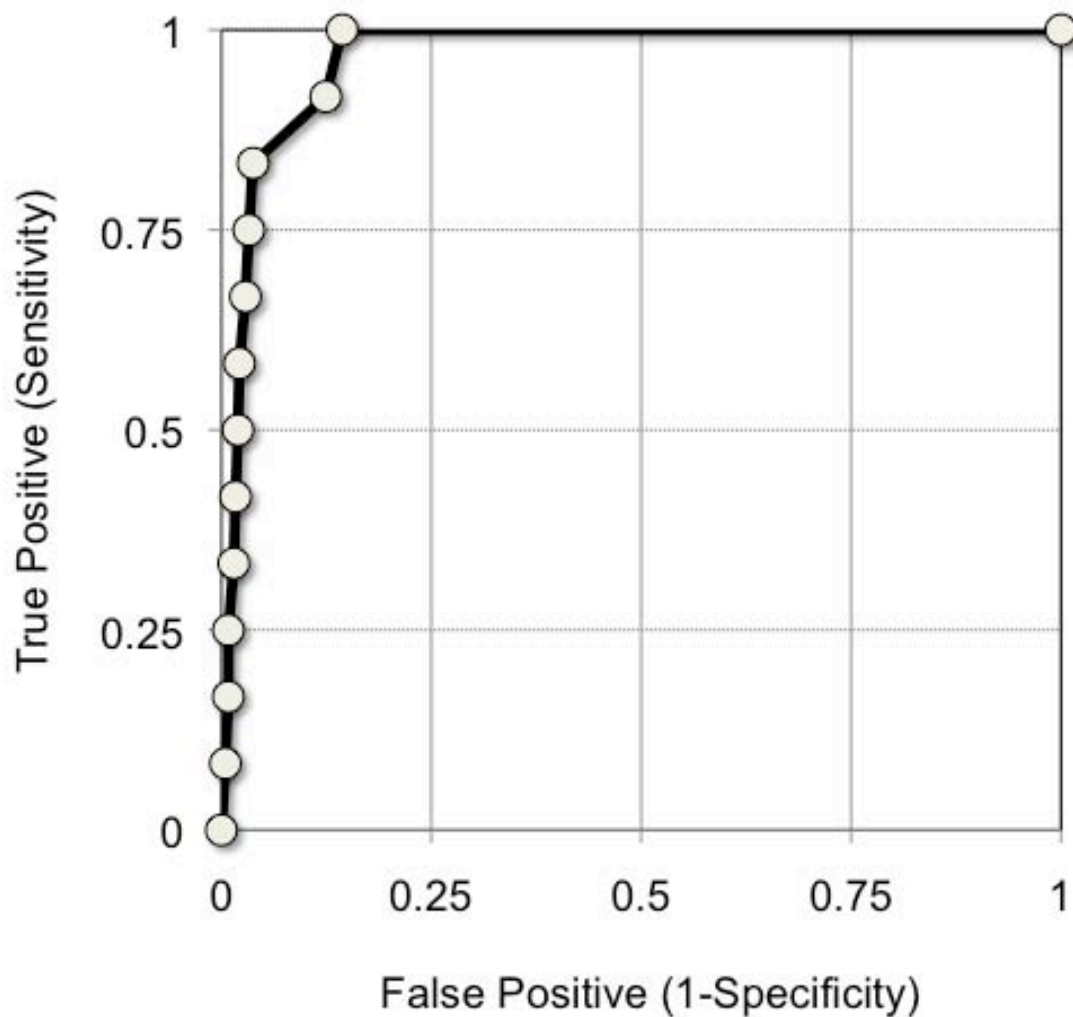
($p=7.6 \times 10^{-11}$, 'Two Locus' model). Removal of *CTLA4* significantly worsened the model ($p=2.1 \times 10^{-7}$, 'Single Locus/*CD28*' model), as did removal of *CD28* ($p=3.4 \times 10^{-3}$, 'Single Locus/*CTLA4*' model).

References

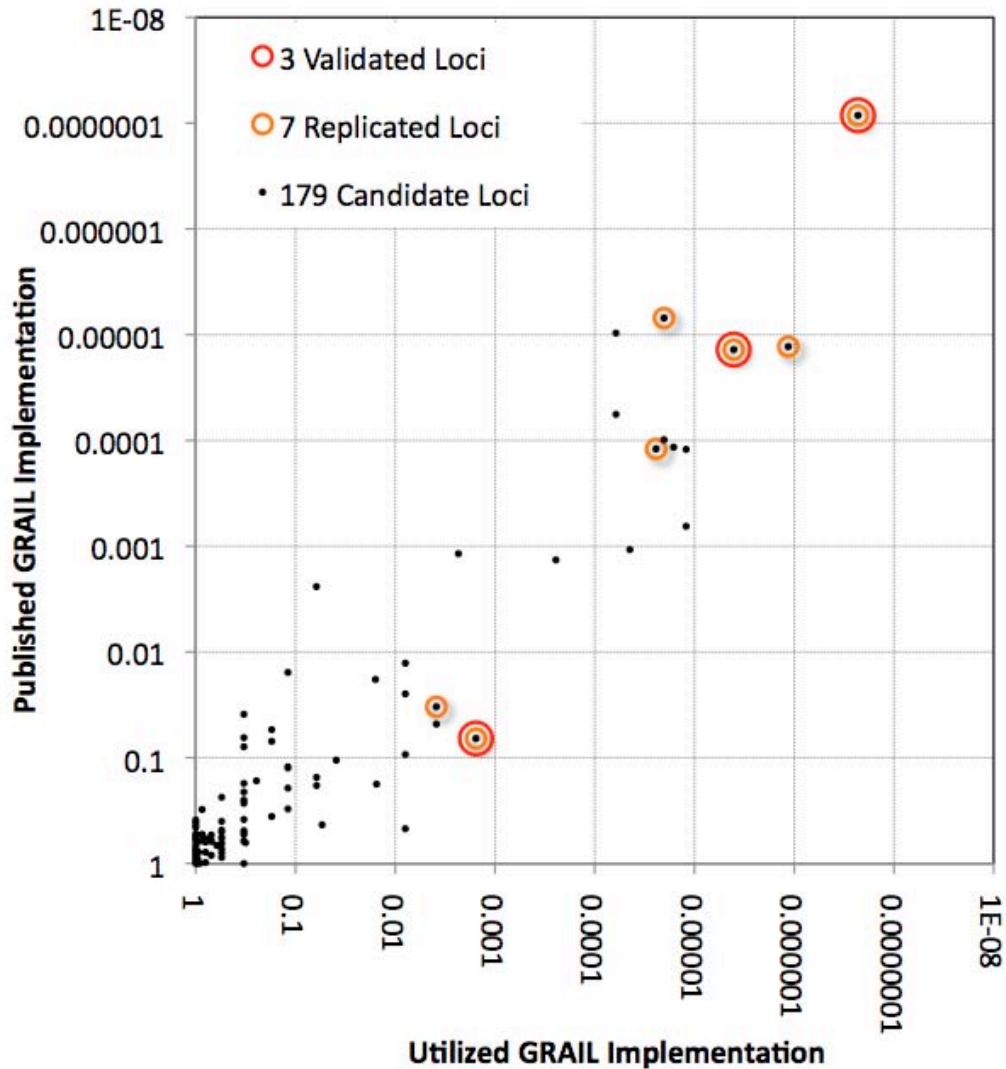
1. Raychaudhuri, S. et al. Common variants at *CD40* and other loci confer risk of rheumatoid arthritis. *Nat Genet* **40**, 1216-23 (2008).
2. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **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- κ B 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_{ext} 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	<i>PTPN22</i>
1p36.13*	rs2240340	0.0012	<i>PADI4</i>
1p36.32	rs3890745	1.00E-06	<i>TNFRSF14</i>
1q23.3	rs6682654	0.0012	<i>CD244</i>
2q33.2*	rs3087243	3.00E-07	<i>CTLA4</i>
2q32.3	rs7574865	1.40E-04	<i>STAT4</i>
4q27	rs6822844	1.10E-05	<i>IL2 IL21</i>
6q23.3	rs10499194, rs6920220	9.10E-05	<i>TNFAIP3</i>
6p21.32 (MHC class II)*	rs6457620	0.0043	<i>HLA-DRA</i>
7q21.2	rs42041	0.25	<i>CDK6</i>
9q33.2	rs3761847	2.30E-04	<i>TRAF1</i>
9p13.3	rs2812378	4.50E-05	<i>CCL21</i>
10p15.1	rs4750316	0.0018	<i>PRKCQ</i>
12q13.3	rs1678542	0.15	<i>B4GALNT1</i>
20q13.12	rs4810485	8.40E-06	<i>CD40</i>
22q12.3	rs3218253	6.10E-04	<i>IL2RB</i>

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

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 al 2008).

Supplementary Table 3.

	Case Collection	Control Collection	Case Origin	Case Antibody Status	Cases	Controls	Genotyping Platform	Case-Controls Stratification
Meta-Analysis 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
	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 7,957 Cases; 11,958 Controls	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
	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	<i>NHLH2</i>	-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	<i>CD2, IGSF2, CD58</i>	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	<i>FCGR2A</i>	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	<i>CD247</i>	-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	<i>PTPRC</i>	-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	<i>TTC7A</i>	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	<i>IL1RL1</i>	-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	<i>CD28</i>	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	<i>PLCL2</i>	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	<i>TNIP2</i>	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	<i>CD109</i>	-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	<i>PRDM1</i>	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	<i>TAGAP</i>	-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	<i>RARRES2</i>	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	<i>IFNB1</i>	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	<i>RAG1, TRAF6</i>	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	<i>CD3G</i>	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	<i>IRAK4</i>	-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	<i>PIBF1</i>	-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	<i>PTPN2</i>	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	<i>ICAM1, ICAM3</i>	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	<i>ICAM1, ICAM3</i>	-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*).