nature genetics

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

Soumya Raychaudhuri[1–3*], Brian P Thomson[2], Elaine F Remmers[4], Stephen Eyre[5], Anne Hinks[5], Candace Guiducci[2], Joseph J Catanese[6], Gang Xie[7], Eli A Stahl[1], Robert Chen[1], Lars Alfredsson[8], Christopher I Amos[9], Kristin G Ardlie[2], BIRAC Consortium[28], Anne Barton[5], John Bowes[5], Noel P Burtt[2], Monica Chang[6], Jonathan Coblyn[1], Karen H Costenbader[1], Lindsey A Criswell[10], J Bart A Crusius[11], Jing Cui[1], Phillip L De Jager[2,12], Bo Ding[8], Paul Emery[13], Edward Flynn[5], Pille Harrison[14], Lynne J Hocking[15], Tom W J Huizinga[16], Daniel L Kastner[4], Xiayi Ke[5], Fina A S Kurreeman[1,16], Annette T Lee[17], Xiangdong Liu[7], Yonghong Li[6], Paul Martin[5], Ann W Morgan[13], Leonid Padyukov[18], David M Reid[15], Mark Seielstad[19], Michael F Seldin[20], Nancy A Shadick[1], Sophia Steer[21], Paul P Tak[22], Wendy Thomson[5], Annette H M van der Helm-van Mil[16], Irene E van der Horst-Bruinsma[23], Michael E Weinblatt[1], Anthony G Wilson[24], Gert Jan Wolbink[25,26], Paul Wordsworth[14], YEAR Consortium[28], David Altshuler[2,3], Elizabeth W Karlson[1], Rene E M Toes[16], Niek de Vries[22], Ann B Begovich[6,27], Katherine A Siminovitch[7], Jane Worthington[5], Lars Klareskog[18], Peter K Gregersen[17], Mark J Daly[2,3] & Robert M Plenge[1,2]

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 controls[1]. 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 polyarthritis[12]. 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-κB 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 methods[14] 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.

Received 13 May; accepted 21 September; published online 8 November 2009; doi:10.1038/ng.479

Table 1 Sixteen validated RA loci used in functional analyses

Validated RA locus	Representative allele (SNPs)	Genes within associated regions
1p13.2[a]	rs2476601	***PTPN22,*** *AP4B1, RSBN1, BCL2L15, DCLRE1B, MAGI3, PHTF1*
1p36.13[a]	rs2240340	*PADI3,* ***PADI4***
1p36.32	rs3890745	*PANK4, MMEL1, PLCH2, HES5,* ***TNFRSF14***
1q23.3	rs6682654	*LY9,* ***CD244***
2q33.2[a]	rs3087243	*ICOS,* ***CTLA4***
2q32.3	rs7574865	*STAT1, GLS,* ***STAT4***
4q27	rs6822844	***IL2, IL21,*** *ADAD1, KIAA1109*
6q23.3	rs10499194, rs6920220	*OLIG3,* ***TNFAIP3***
6p21.32[a] (MHC class II)	rs6457620, DRB1*0401, *0101	***HLA-DRA,*** *HLA-DQB1, BTNL2, HLA-DQA1, HLA-DRB5, HLA-DRB1*
7q21.2	rs42041	*PEX1, FAM133B, GATAD1,* ***CDK6***
9q33.2	rs3761847	*PHF19, CEP110,* ***TRAF1,*** *RAB14, C5*
9p13.3	rs2812378	***CCL21***
10p15.1	rs4750316	*RBM17, PFKFB3,* ***PRKCQ***
12q13.3	rs1678542	*DTX3, METTL1, AVIL, DDIT3, XRCC6BP1, MBD6, GLI1, CYP27B1, KIF5A, GEFT, CTDSP2, MARS, CDK4, AGAP, DCTN2, TSPAN31, FAM119B, MARCH9, TSFM,* ***B4GALNT1,*** *OS9, PIP4K2C, ARHGAP9, SLC26A10*
20q13.12	rs4810485	*SLC12A5, NCOA5,* ***CD40***
22q12.3	rs3218253	***IL2RB***

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.
[a]Loci 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 $P_{text} < 0.01$. 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 $P_{text}<0.01$; 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 controls[1]. 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 $P_{text} < 0.01$ based on the estimated 5.4% false positive rate. However, 22 of the 179 (12.3%) scored $P_{text} < 0.01$ (**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

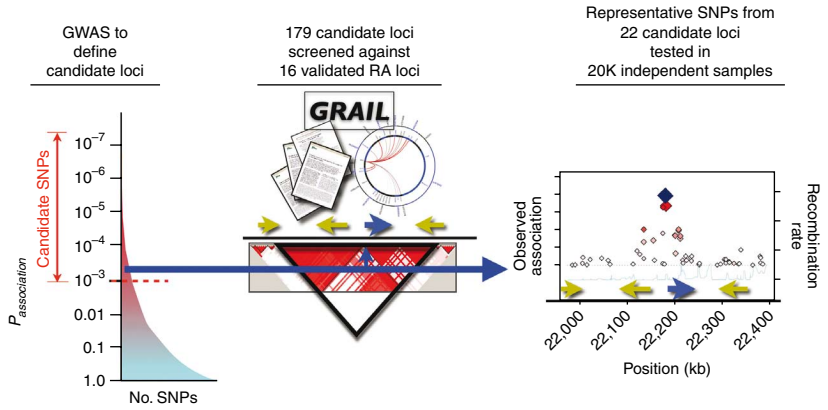

Figure 1 Using Gene Relationships Across Implicated Loci (GRAIL) to prioritize candidate RA SNPs. We selected a set of candidate SNPs to pursue in an independent genotyping experiment by starting with all SNPs that obtained $P < 0.001$ in an independent GWAS meta-analysis. Then, for each candidate SNP, 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.

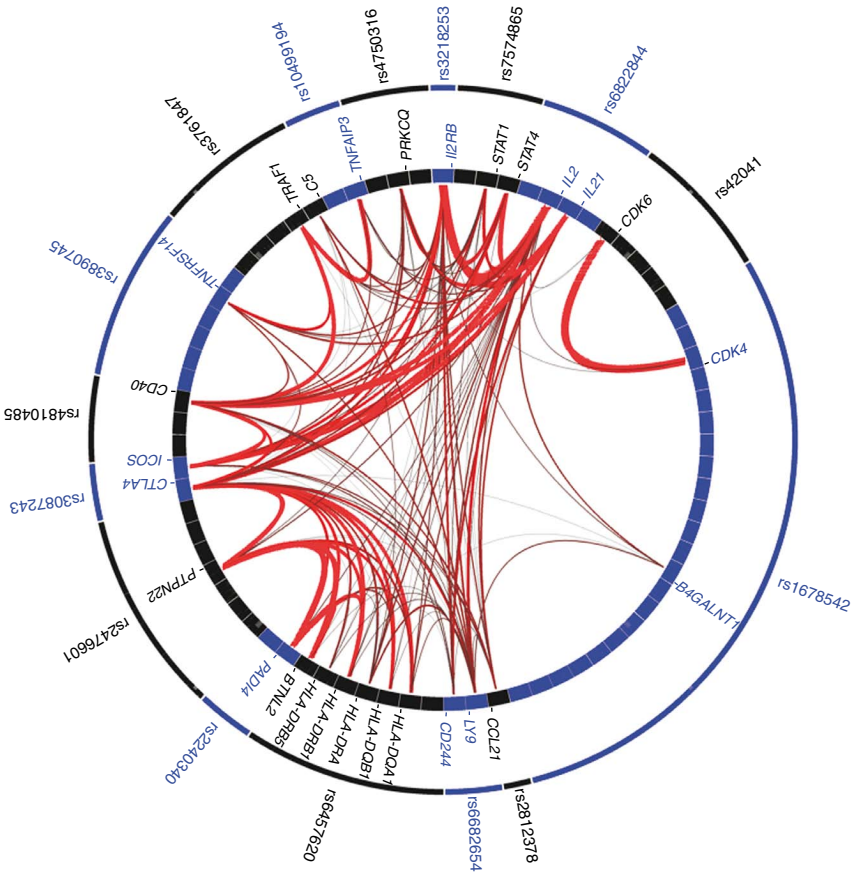

Figure 2 GRAIL identified interconnectivity among genes in RA loci. The known RA-associated SNPs are along the outer ring; the internal ring represents the genes near each SNP (as listed in **Table 1**); each box in the internal ring represents an individual gene. We illustrate the literature-based functional 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 criteria[15] 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 statistic[16]. 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 $P_{skew} = 0.0005$, 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

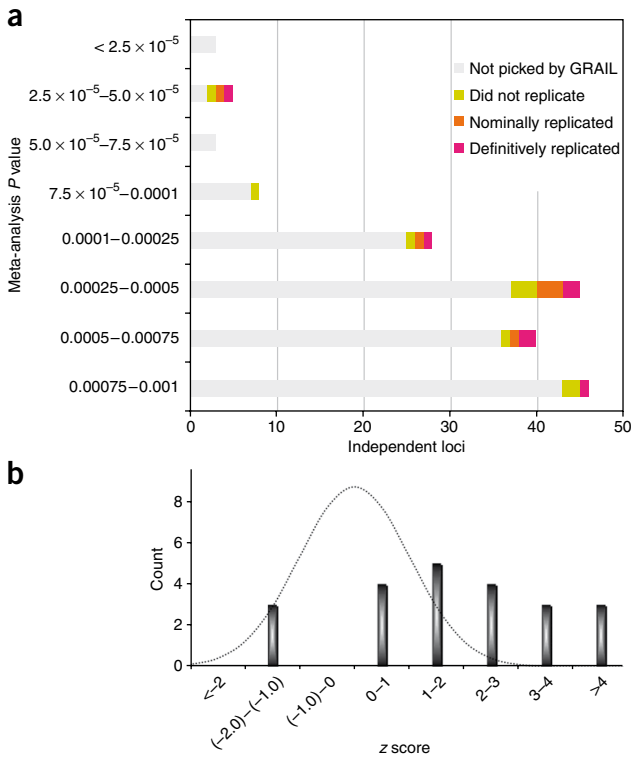

Figure 3 GRAIL identifies new RA risk loci that replicate when genotyped in independent case-control samples. (**a**) GRAIL identified 22 SNPs among the 179 candidate SNPs with $P < 0.001$ in a GWAS meta-analysis. This is a histogram of the 179 SNPs as a function of their GWAS meta-analysis P value. 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

SNP						Meta-analysis				Replication				Joint		
				Allele				Minor allele				Minor allele				Breslow-Day
ID	Chr	Pos (HG18)	Gene(s)	Major	Minor	*P*	OR	Control	Case	*P*	OR	Control	Case	*P*	OR	*P*
Replicated loci (uncorrected $P < 0.0023$)																
rs11586238[a]	1p13.1	117,064,661	*CD2, IGSF2, CD58*	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	*CD28*	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	*PRDM1*	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	*TAGAP*	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	*PTPRC*	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	*RAG1, TRAF6*	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	*FCGR2A*	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	*PTPN2*	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	*PLCL2*	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	*CD247*	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	*ICAM1, ICAM3*	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	*NHLH2*	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	*TNIP2*	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	*CD3G*	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	*ICAM1, ICAM3*	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	*RARRES2*	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	*IFNB1*	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	*PIBF1*	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	*TTC7A*	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	*IL1RL1*	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	*IRAK4*	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	*CD109*	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.
[a]These 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 sclerosis[22–24]. 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 disease[20]; 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	*FCGR2A*	rs1801274	*FCGR2A*	Systemic lupus erythematosus	12.7	0.19	1.00
rs394581	6q25.3	*TAGAP*	rs1738074	*TAGAP*	Celiac disease, type 1 diabetes	16.5	0.32	0.73
rs10919563	1q31.3	*PTPRC*	rs17612648	*PTPRC*	Multiple sclerosis	34.5	—	—
rs7234029	18p11.21	*PTPN2*	rs478582	*PTPN2*	Type 1 diabetes	41.1	0.14	1.00
rs1980422	2q33.2	*CD28*	rs3087243	*CTLA4*	Type 1 diabetes, RA	128.5	0.04	0.40
rs548234	6q21	*PRDM1*	rs7746082	*PRDM1*	Crohn's disease	132.8	0.01	0.08
rs11586238	1p13.1	*CD2*	rs2300747	*CD58*	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 disease[25]. 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-κB 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 directly[28]. SNP association to CD28 contributes additional evidence of the role of T-cell activation in disease pathogenesis. TRAF6 is involved in downstream NF-κB activation; it binds CD40 directly and is a key component of B-cell activation[29]. 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 cells[30]. 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-κB 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).

[1]Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, USA. [2]Broad Institute, Cambridge, Massachusetts, USA. [3]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. [4]Genetics and Genomics Branch, National Institute of Arthritis and Musculoskeletal and Skin Diseases, US National Institutes of Health, Bethesda, Maryland, USA. [5]Arthritis Research Campaign (arc)–Epidemiology Unit, The University of Manchester, Manchester, United Kingdom. [6]Celera, Alameda, California, USA. [7]Department of Medicine, University of Toronto, Mount Sinai Hospital and University Health Network, Toronto, Ontario, Canada. [8]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [9]University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. [10]Rosalind Russell Medical Research Center for Arthritis, Department of Medicine, University of California, San Francisco, California, USA. [11]Laboratory of Immunogenetics, Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands. [12]Department of Neurology, Center for Neurologic Diseases, Brigham and Women's Hospital, Boston, Massachusetts, USA. [13]National Institute for Health Research-Leeds Musculoskeletal Biomedical Research Unit, Leeds Institute of Molecular Medicine, University of Leeds, United Kingdom. [14]University of Oxford Institute of Musculoskeletal Sciences, Botnar Research Centre, Oxford, United Kingdom. [15]Musculoskeletal and Genetics Section, Division of Applied Medicine, University of Aberdeen, United Kingdom. [16]Department of Rheumatology, Leiden University Medical Centre, Leiden, The Netherlands. [17]The Feinstein Institute for Medical Research, North Shore-Long Island Jewish Health System, Manhasset, New York, USA. [18]Rheumatology Unit, Department of Medicine, Karolinska Institutet at Karolinska University Hospital Solna, Stockholm, Sweden. [19]Genome Institute of Singapore, Singapore. [20]Rowe Program in Genetics, University of California at Davis, Davis, California, USA. [21]Clinical and Academic Rheumatology, Kings College Hospital National Health Service Foundation Trust, Denmark Hill, London, United Kingdom. [22]Clinical Immunology and Rheumatology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. [23]Department of Rheumatology, VU University Medical Center, Amsterdam, The Netherlands. [24]School of Medicine and Biomedical Sciences, Sheffield University, Sheffield, United Kingdom. [25]Jan van Breemen Institute, Amsterdam, The Netherlands. [26]Sanquin Research Landsteiner Laboratory, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. [27]Roche Diagnostics, Pleasanton, California, USA. [28]A 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 loci[2]. 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 $P_{text} < 0.01$ significance score; the specificity was defined as the proportion of spurious associations that GRAIL assigned a $P_{text} > 0.01$ 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 study[1]. 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 ($P_{text} < 0.01$) 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 implementation[2]—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 sclerosis[32], age-related macular degeneration (B.M. Neale, J. Fagerness, R. Reynolds, L. Sobrin, M. Parker, S. Raychaudhuri *et al.* unpublished results) and myocardial infarction[33]; (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 cancer[34]; (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 study[40] 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 use[41]; and (xi) CCP- or RF-positive cases recruited at multiple sites in the United Kingdom by the United Kingdom Rheumatoid Arthritis Genetics (UKRAG) collaboration[6]. 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 PCR[42] (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 EIGENSTRAT[43]. 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 elsewhere[1].

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. Nielen, 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 TNFα 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).

doi:10.1038/ng.479