Two independent alleles at 6q23 associated with risk of rheumatoid arthritis

Robert M Plenge1–3, Chris Cotsapas1,3, Leela Davies1, Alkes L Price1,4, Paul I W de Bakker1,3,4, Julian Maller1,3, Itsik Pe’er3, Noel P Burtt1, Brendan Blumenstiel1, Matt DeFelice1, Melissa Parkin1, Rachel Barry1, Wendy Winslow1, Claire Healy1, Robert R Graham1,3, Benjamin M Neale1,3,6, Elena Izmailova7, Ronenn Roubenoff7, Alexander N Parker7, Roberta Glass7, Elizabeth W Karlson2, Nancy Maher2, David A Hafler1,8, David M Lee2, Michael F Seldin9, Elaine F Remmers10, Annette T Lee11, Leonid Padyukov12, Lars Alfredsson12, Jonathan Coblyn2, Michael E Weinblatt2, Stacey B Gabriel1, Shaun Purcell1,3, Lars Klareskog12, Peter K Gregersen11, Nancy A Shadick2, Mark J Daly1,3 & David Altshuler1,3,4

To identify susceptibility alleles associated with rheumatoid arthritis, we genotyped 397 individuals with rheumatoid arthritis for 116,204 SNPs and carried out an association analysis in comparison to publicly available genotype data for 1,211 related individuals from the Framingham Heart Study1. After evaluating and adjusting for technical and population biases, we identified a SNP at 6q23 (rs10499194, ~150 kb from TNFAIP3 and OLIG3) that was reproducibly associated with rheumatoid arthritis both in the genome-wide association (GWA) scan and in 5,541 additional case-control samples (P = 10−3, GWA scan; P < 10−6, replication; P = 10−9, combined). In a concurrent study, the Wellcome Trust Case Control Consortium (WTCCC) has reported strong association of rheumatoid arthritis susceptibility to a different SNP located 3.8 kb from rs10499194 (rs6920220; P = 5 × 10−6 in WTCCC)2. We show that these two SNP associations are statistically independent, are each reproducible in the comparison of our data and WTCCC data, and define risk and protective haplotypes for rheumatoid arthritis at 6q23.

Rheumatoid arthritis is the most common inflammatory arthritis, affecting up to 1% of the adult population3. Two loci (HLA-DRBI4 and PTPN225) have previously been associated with rheumatoid arthritis susceptibility in individuals with circulating antibodies to cyclic citrullinated peptides (CCP). Most of the inheritance of rheumatoid arthritis remains unexplained.

To identify additional common variants associated with risk of CCP antibody–associated (CCP+) rheumatoid arthritis, we conducted a GWA study using the Affymetrix 100K GeneChip microarray in a longitudinal case series of individuals with CCP+ rheumatoid arthritis (the Brigham Rheumatoid Arthritis Sequential Study (BRASS) cohort). As we lacked epidemiologically matched controls, we compared case data to publicly available genotype data collected using the same platform from 1,211 related Framingham Heart Study (FHS) participants1, drawn from the same geographical region as the individuals in our study (near Boston, Massachusetts, USA).

Before comparing allele frequencies between cases and controls, we considered biases that may be introduced by the use of shared controls. Such biases, whether due to nonrandom distribution of technical artifacts6 or to population differences between case and control data7,8, would result in a non-null distribution of test statistics with excess false-positive associations. In an initial analysis of unrelated case-control samples, we assessed the median distribution of test statistics with the genomic-control parameter λGC9 (where 1.0 indicates no inflation) and examined the tail of the distribution of association statistics in a comparison of observed and expected P values (Q-Q plot; Fig. 1).

Using published data quality control parameters from early studies on this genotyping platform (genotype call rates > 90%, minor allele...
frequency (MAF) > 5%)\textsuperscript{1}, we observed $\lambda_{GC} = 1.19$ and an excess of associations in the extreme tail of the $-\log_{10}(P)$ distribution (Fig. 1a).

To disentangle the contribution of genotyping bias from that due to population stratification, we examined the $\chi^2$ distribution for a subset of 40,562 SNPs with nearly complete genotype data (call rate > 99%). This stringent filtering of SNPs reduced $\lambda_{GC}$ to 1.12, and fewer SNPs had extreme $P$ values (Fig. 1b and Supplementary Table 1 online), indicating that SNPs with low call rates were disproportionately inflating the association statistics. The presence of residual inflation in the $\chi^2$ distribution, however, suggested that bias in missing genotype data was not the only source of inflation in this study.

We next used two statistical methods to adjust for inflation due to population stratification: structured association by genetically matching cases and controls using identity-by-state similarity as implemented in PLINK\textsuperscript{10} and a principal components approach (EIGENSTRAT)\textsuperscript{11}. After these adjustments, $\lambda_{GC}$ was nearly completely normalized, falling from 1.12 to 1.04 (PLINK Cochrane-Mantel-Haenszel; Fig. 1c) and 1.03 (EIGENSTRAT; Supplementary Table 1), with both methods giving very similar results (Supplementary Fig. 1 online). Thus, using a set of SNPs with complete genotype data and controlling for stratification in either of two ways, we found that an essentially null distribution of association statistics could be obtained despite the use of shared controls and a first-generation genotyping platform with substantial missing data.

Although this approach accounted for observed biases, it did so at the cost of reduced genome coverage due to stringent SNP filtering: from 30% of common HapMap CEU SNPs captured (at $r^2 > 0.8$) by the 87,962 SNPs with call rates > 90% to just 18% captured with the subset of 40,562 with call rates > 99%. In a two-parameter linear model with call rate and minor-allele frequency as variables, we found that $\lambda_{GC}$ was considerably associated with call rate and with an interaction between call rate and MAF (Supplementary Fig. 2 online). Thus, instead of a standard correction of uniformly dividing all test statistics by $\lambda_{GC}$ we used linear regression to correct the test statistics of 79,853 SNPs with > 95% call rates as a function of call rate and MAF–call rate interaction (Supplementary Fig. 3 online). This dynamic genomic-control correction resulted in a null $-\log_{10}(P)$ distribution (Fig. 1d) and maintained genome coverage at 29% of HapMap CEU SNPs.

Finally, as the available control genotypes were drawn from related individuals from multigenerational pedigrees, we evaluated whether power was improved by including genotypes from multiple related individuals (adjusting for the inflation in the $\chi^2$ distribution) or by using only the unrelated individuals from each pedigree (see Supplementary Methods and Supplementary Fig. 4 online). Specifically, we evaluated significance for the two known true-positive associations (HLA-DRB1 and PTPN22) in each design. Inclusion of related individuals predictably inflated the $\chi^2$ distribution, with $\lambda_{GC}$ increasing from 1.04 to 1.34 (Supplementary Table 2 online) because of overestimation of the number of control chromosomes (as some are not independent). However, even after correction for this inflation, we observed a net increase in ability to detect the effect of HLA-DRB1 and PTPN22 (Supplementary Table 2). Intuitively, this is not surprising, as inclusion of additional family members increases the number of independent chromosomes with which to estimate control-allele frequencies.

On the basis of these evaluations, we carried out association analysis of 397 CCP+ rheumatoid arthritis cases and 1,211 related FHS controls over 79,853 SNPs, using PLINK CMH to correct for stratification, two-parameter linear modeling to correct for genotype artifact, and residual $\lambda_{GC}$ to correct for relatedness. This analysis resulted in an overall null distribution of results, with only slight enrichment in the tail, where an excess of spurious results may have occurred (Fig. 1e).

Such enrichment could be due to true-positive results, or it could be due to bias that we failed to account for in our study. We report...
Table 1 Summary of results for rs10499194 across 2,680 CCP+ rheumatoid arthritis cases and 4,469 controls

<table>
<thead>
<tr>
<th>Collection</th>
<th>n (case)</th>
<th>n (control)</th>
<th>(\lambda_{GC}) SNPs</th>
<th>(\lambda_{GC})</th>
<th>(P) value (corr)</th>
<th>(\lambda_{GC})</th>
<th>(P) value (corr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRASS versus FHS</td>
<td>397</td>
<td>1,211</td>
<td>80K panel</td>
<td>1.34</td>
<td>0.0009 (0.001)</td>
<td>n.a.</td>
<td>0.39*</td>
</tr>
<tr>
<td>EIRA</td>
<td>875</td>
<td>832</td>
<td>n.a.</td>
<td>n.a.</td>
<td>0.0003 (0.0004)</td>
<td>n.a.</td>
<td>0.39*</td>
</tr>
<tr>
<td>NARAC (family)</td>
<td>535</td>
<td>1,013</td>
<td>704 AIMs</td>
<td>1.33</td>
<td>0.0000008 (0.0007)</td>
<td>1.30</td>
<td>0.000004 (0.0005)</td>
</tr>
<tr>
<td>NARAC (sporadic)</td>
<td>873</td>
<td>1,413</td>
<td>704 AIMs</td>
<td>2.70</td>
<td>0.000002 (0.01)</td>
<td>1.28</td>
<td>0.006 (0.02)</td>
</tr>
<tr>
<td>Total</td>
<td>2,680</td>
<td>4,469</td>
<td></td>
<td>6 \times 10^{-12} (2 \times 10^{-8})</td>
<td>1 \times 10^{-9} (3 \times 10^{-8})</td>
<td>0.75 (0.66-0.87)</td>
<td></td>
</tr>
</tbody>
</table>

Two-tailed \(P\) values are shown for PLINK CMH and EIGENSTRAT, where either the 80K SNP panel or 704 AIM SNPs was used to correct for population stratification and calculate residual inflation with \(\lambda_{GC}\), as indicated. The asterisks (*) next to the \(P\) values for EIRA indicate that these were calculated using \(2 \times 2\) contingency tables of allele frequencies using a standard \(\chi^2\) test. In BRASS and NARAC (family and sporadic collections), we provide an additional correction for residual inflation with \(\lambda_{GC}\) (corr). The additional correction-based \(\lambda_{GC}\) calculated with AIM SNPs are very conservative, as these SNPs were selected to differentiate Northern versus Southern European ancestry, and as such will overestimate the amount of inflation compared to a randomly selected set of SNPs. (In NARAC, for example, residual \(\lambda_{GC}\) after EIGENSTRAT is 1.03 for the 21 replication SNPs.) In EIRA, no additional genotype data were available to apply methods to correct for stratification. The final combined \(P\) value we report in the abstract and text is based on Fisher's method of combining \(P\) values using EIGENSTRAT to correct for stratification in the original GWA scan and in the NARAC replication samples (\(P = 10^{-12}\)). A combined odds ratio was generated using a random effects model. n.a., not applicable.

Before learning of the WTCCC results, in an attempt to fine map our association, we had genotyped in our replication samples an additional 17 SNPs chosen on the basis of imperfect linkage disequilibrium (LD) to rs10499194 (\(r^2 = 0.20-0.95\)). In light of the WTCCC results, we carried out stepwise regression analysis to determine whether the two signals were independent or simply due to linkage disequilibrium with each other or another SNP in the region. Specifically, we used these 17 SNPs to predict SNPs in CEU HapMap individuals that were not directly genotyped in our study but that could be well predicted using single SNPs or multi-marker haplotypes. In this analysis, the SNP we originally observed (rs10499194) provided a strong signal of association (Fig. 2) but alone did not explain the entire association signal: the SNP with the stronger association in WTCCC (rs6922020, imputed with \(r^2 = 1\) using a two-marker predictor) remained significant after analysis conditional on the original GWA association results and linkage disequilibrium (LD) structure at 6q23. Results for SNPs genotyped across 1 Mb as part of the original GWA scan in 397 CCP+ rheumatoid arthritis cases and 1,111 related controls (gray diamonds), as well 17 SNPs genotyped in additional replication samples (2,283 unrelated CCP+ rheumatoid arthritis cases and 3,258 unrelated controls). In the replication samples, the color of each diamond is based on \(r^2\) (CEU HapMap) with the most significant SNP in our study (rs10499194). The blue diamond indicates the \(P\) value for all samples in our study (original GWA scan plus replication samples), as determined by Fisher’s method of combining \(P\) values (EIGENSTRAT for the original GWA scan and replication samples). The recombination rate based on CEU HapMap is shown in light blue along the x-axis (scale on the right); the red line indicates a 63-kb region of strong LD used to construct haplotypes. The green arrows indicate gene location; the associated SNP is \(\sim 185\) kb away from either TNFAIP3 or OLG3.
on rs10499194 (P = 0.0005 for rs6920220; MAF = 0.241 for cases and 0.196 for controls). Analysis of rs6920220 alone was also highly significant (P = 2.8 × 10−12). Addition of other SNPs to the haplotype analysis did not increase the significance of the model, and the two SNPs together did not predict any known HapMap SNP. These two SNPs reside on distinct phylogenetic branches of the haplotype tree constructed with genotype data from our study and define three categories of risk: a ‘protective’ haplotype tagged by rs10499194; a ‘risk’ haplotype tagged by rs6920220; and the remaining haplotypes, which have risks equal to one another (Fig. 3). Although these data strongly suggest the existence of two independent susceptibility alleles, exhaustive resequencing is required to rule out the possibility that these two SNPs form a haplotype in LD with a single, as-yet-undefined causal allele. If multiple independent association signals are confirmed, the finding of multiple common risk alleles at 6q23 would be similar to other recent examples of multiple alleles such as the associations of IRF5 and risk of systemic lupus erythematosus15, IL23R and risk of Crohn’s disease16, 8q24 and risk of prostate cancer17–19 and CFH and risk of age-related macular degeneration20.

These two SNPs (rs10499194 and rs6920220) are located within 3.8 kb of each other but are >150 kb from the nearest genes, which are those encoding tumor necrosis factor, alpha-induced protein 3 (TNFAIP3, ~185 kb telomeric), and oligodendrocyte transcription factor 3 (OLIG3, ~185 kb centromeric; Fig. 2). TNFAIP3, also known as A20, is a potent inhibitor of NF-κB signaling and is required for termination of tumor necrosis factor (TNF)-induced signals21. TNF-α levels are increased in individuals with rheumatoid arthritis, and inhibition of TNF-α is a potent treatment of severe rheumatoid arthritis22. Furthermore, mice lacking Tnfaip3 show chronic inflammation23, consistent with loss of function of this gene playing a role in autoimmunity. Far less is known about OLIG3. Mutant Olig3 mice have abnormalities in neuronal development but no reported abnormalities of the immune or musculoskeletal systems24. Finally, two other immune-related genes lie within 1 Mb of the associated region (IL22RA and IFNGR1). Additional genetic and functional studies will be required to determine which of these genes, or others not yet recognized, explain the two SNP associations observed consistently and significantly across our study and the WTCCC results.

**METHODS**

**BRASS rheumatoid arthritis cases and FHS control samples.** Samples from patients with rheumatoid arthritis (n = 435) were collected at Brigham and Women’s Hospital in Boston, Massachusetts (USA), as part of the BRASS Registry25. A total of 1,343 Framingham Heart Study samples from 303 multiplex families were available for analysis. Because the population prevalence of rheumatoid arthritis is <1% in the adult population, and because only limited data on the rheumatoid arthritis status of FHS samples were available, all FHS samples were considered as possible controls. Informed consent was obtained by the institutions overseeing the BRASS and FHS studies.

Affymetrix SNP genotyping and initial quality-control filtering. Genotyping of the rheumatoid arthritis samples was carried out at the Broad Institute using the Affymetrix GeneChip 100K Mapping Array containing 116,204 SNPs. FHS samples were genotyped at Boston University1 and obtained through a formal application process. Genotypes were called using the dynamic-modeling algorithm. (BRLMM data were available for the rheumatoid arthritis samples, but we did not use them because we only had access to FHS genotypes called using the dynamic-modeling algorithm.) Both datasets were filtered individually and then merged; individuals with >10% missing genotypes and SNPs with >10% missing data or Hardy-Weinberg equilibrium (HWE) P values <0.0001 were excluded. After applying these filters, 405 rheumatoid arthritis cases and 1,305 FHS controls remained. We removed FHS individuals with two genotyped parents (n = 66), as these samples contribute no independent genetic information. The average call rate of the 87,962 SNPs across these samples was 98.3%. The rheumatoid arthritis–associated SNP (rs10499194) had a call rate of 98.03% in the rheumatoid arthritis cases and 99.24% in FHS controls, with a HWE P value >0.05. Additional details are available in Supplementary Methods. The Massachusetts Institute of Technology Institutional Review Board approved the study.
GWA study using PLINK and EIGENSTRAT. We compared SNP allele frequency in unrelated rheumatoid arthritis samples to either unrelated (n = 393) or related (n = 1,211) FHS controls. In analysis without correction for population stratification, significance was determined using standard Pearson’s χ² test for contingency tables. To correct for population stratification, we first removed genetic outliers (see Supplementary Methods) and then applied two distinct methods: Cochrane-Mantel-Haenszel (CMH) meta-analysis implemented in PLINK20 and a principal-components method implemented in EIGENSTRAT21. We used PLINK CMH for our primary analysis and EIGENSTRAT for a secondary analysis (Supplementary Methods).

Linear model (dynamic genomic control) correction. We first normalized the distribution of association statistics by taking the square root and arbitrarily changing sign for SNPs whose odds ratios were >1. This resulted in an essentially normal distribution of values, to which we fit a linear model with two parameters: missing data proportion and minor allele frequency, including their interaction. Corrected test statistics were recovered by inverting the normalization process for residuals of the model.

Replication samples. Our overall strategy was to replicate our top SNPs in two sample collections: population-based case-control samples from Sweden (EIRA12) and familial case-control samples from North America (NARAC sample collections: population-based case-control samples from Sweden and Utah, and replication samples. For NARAC, our primary analysis was EIGENSTRAT11 applied to a set of 704 AIMs and unrelated FHS samples and 99.8% for the related FHS samples. Additional genotype data of 704 European ancestry informative markers (AIMs) had been previously carried using the Illumina GoldenGate custom assay27 and were available in all NARAC samples.

Statistical analysis of rs10499194 in replication data. Our primary analysis in EIRA was based on 2 × 2 contingency tables of allele frequencies and a χ² test. For NARAC, our primary analysis was EIGENSTRAT11 applied to a set of 704 European substructure AIMs27 and correcting along the first principal component. As a secondary analysis in NARAC, we used the 704 AIMs to generate identity-by-state case-control clusters (for CMH analysis in PLINK; see Supplementary Methods).

Statistical analysis of additional SNPs and haplotypes in replication data. We combined replication genotype data for all 2,283 unrelated CCDP+ rheumatoid arthritis cases and 3,258 unrelated control samples were available for analysis. We received permission from FHS to genotype a single SNP rs10499194, in the same set of FHS samples. The Affymetrix-Seqeuom concordance for rs10499194 was 100% for the BRASS and unrelated FHS samples and 99.8% for the related FHS samples. Additional genotype data of 704 European ancestry informative markers (AIMs) had been previously carried using the Illumina GoldenGate custom assay27 and were available in all NARAC samples.

Statistical analysis of rs10499194 in replication data. Our primary analysis in EIRA was based on 2 × 2 contingency tables of allele frequencies and a χ² test. For NARAC, our primary analysis was EIGENSTRAT11 applied to a set of 704 European substructure AIMs27 and correcting along the first principal component. As a secondary analysis in NARAC, we used the 704 AIMs to generate identity-by-state case-control clusters (for CMH analysis in PLINK; see Supplementary Methods).

Competing interests statement
The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Author contributions

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

Acknowledgments
The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University or NHLBI.

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

References

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics