

ORIGINAL ARTICLE

High-density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33

CI Amos¹, WV Chen¹, A Lee², W Li², M Kern², R Lundsten², F Batliwalla², M Wener³, E Remmers⁴, DA Kastner⁴, LA Criswell⁵, MF Seldin⁶ and PK Gregersen²

¹Department of Epidemiology, University of Texas, MD Anderson Cancer Center, Houston, TX, USA; ²Robert S Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, Manhasset, NY, USA; ³Department of Laboratory Medicine, University of Washington School of Medicine, Seattle, USA; ⁴Genetics and Genomics Branch, NIAMS, NIH, Bethesda, MD, USA; ⁵Department of Medicine, Rosalind Russell Medical Research Center for Arthritis, University of California at San Francisco, San Francisco, CA, USA and ⁶Rowe Program in Genetics, University of California at Davis, Davis, CA, USA

We have completed a genome wide linkage scan using > 5700 informative single-nucleotide polymorphism (SNP) markers (Illumina IV SNP linkage panel) in 642 Caucasian families containing affected sibling pairs with rheumatoid arthritis (RA), ascertained by the North American Rheumatoid Arthritis Consortium. The results show striking new evidence of linkage at chromosomes 2q33 and 11p12 with logarithm of odds (LOD) scores of 3.52 and 3.09, respectively. In addition to a strong and broad linkage interval surrounding the major histocompatibility complex (LOD > 16), regions with LOD > 2.5 were observed on chromosomes 5 and 10. Additional linkage evidence (LOD scores between 1.46 and 2.35) was also observed on chromosomes 4, 7, 12, 16 and 18. This new evidence for multiple regions of genetic linkage is partly explained by the significantly increased information content of the Illumina IV SNP linkage panel (75.6%) compared with a standard microsatellite linkage panel utilized previously (mean 52.6%). Stratified analyses according to whether or not the sibling pair members showed elevated anticyclic citrullinated peptide titers indicates significant variation in evidence for linkage among strata on chromosomes 4, 5, 6 and 7. Overall, these new linkage data should reinvigorate efforts to utilize positional information to identify susceptibility genes for RA. *Genes and Immunity* (2006) 7, 277–286. doi:10.1038/sj.gene.6364295; published online 4 May 2006

Keywords: genetic linkage analysis; single-nucleotide polymorphisms; rheumatoid arthritis; stratified analysis

Introduction

Rheumatoid arthritis (RA) is the most common inflammatory polyarthritis (RA [MIM 180300]). The cause is unknown, although the presence of autoantibodies such as anti-cyclic citrullinated peptide (anti-CCP) autoantibodies, and an association with human lymphocyte antigen (HLA) have led to its categorization as an autoimmune disorder.¹ A genetic component to RA susceptibility has long been established by data from twin and family studies.² Twin analysis, for example, estimated the heritability of RA to be about 60%.³ The strength of the genetic component has also been estimated by computing the relative recurrence risk for siblings (λ_s) of RA probands compared with that for the general population.⁴ Due to uncertainty about the

prevalence of RA in Caucasian populations, as well as variability in the phenotype, the estimated increase in risk to siblings varies between 5 and 10.²

The genetic basis of RA is complex.⁵ Consistent associations clearly implicate a role of the major histocompatibility complex (MHC) in risk for RA.⁶ Indeed, based on previous linkage results, the MHC region makes the largest single contribution ($\lambda_s \sim 1.8$) to disease susceptibility.^{7,8} A set of alleles at the DRB1 locus, many of which share a common polymorphic sequence, the 'shared epitope' (SE),⁹ explain a large portion, but not all,^{10–12} of the genetic risk within the MHC. Recently, the R620W variant of the PTPN22 locus has been shown to confer increased risk for RA, with odds ratios ranging between 1.5 and 2.0 for heterozygotes, and over 3.0 for homozygous carriers of the variant.^{13,14} This finding has been extensively replicated,^{15–17} and is now accepted as the most robust genetic association with RA outside of the MHC.¹⁸ Interestingly, the 620W PTPN22 allele is also associated with several other autoimmune disorders including type 1 diabetes,^{19,20} autoimmune thyroid disease,²¹ systemic lupus erythematosus^{22,23} and some forms of juvenile arthritis.^{24,25} A recent report indicates that additional variability in the PTPN22 locus may account for a more minor proportion of risk for

Correspondence: Dr C Amos, Department of Epidemiology, University of Texas, MD Anderson Cancer Center, 1155 Pressler, Unit 1340, Houston, TX 77030, USA. E-mail: camos@mdanderson.org or PK Gregersen, Feinstein Institute for Medical Research, North Shore LIJ Health System, 350 Community Drive, Manhasset, NY 11030, USA.

E-mail: peterg@nshs.edu

Received 8 December 2005; revised 31 January 2006; accepted 31 January 2006; published online 4 May 2006

RA.²⁶ Aside from PTPN22 and the HLA region, associations with CTLA4 and PADI4 have also been supported in some additional studies,²⁷ suggesting that these genes may also contribute to RA susceptibility in Caucasian populations, albeit with rather modest relative risks.

We and others have previously reported on the results of microsatellite linkage scans on Caucasian affected sibling pair families with RA.^{7,8,28–30} Outside of the MHC, no genetic region has been identified that meets accepted criteria for 'significant' linkage (logarithm of odds (LOD) 3.6 for sibling pair studies³¹). A meta-analytical study³² of data published by groups in the UK, Europe, the US and Japan provided evidence ($P < 0.01$) for loci influencing RA risk on chromosomes 6p, 6q, centromeric 16q and 12p. A relatively small linkage study of RA sibling pairs with over 10 000 SNP markers found moderately increased evidence for linkage outside the MHC²⁹ compared with results that were obtained using a microsatellite scan.

We have now undertaken a genome wide SNP linkage scan on 642 affected sibling pair families, the largest single linkage study in RA reported to date. We utilized the most recent version (IV) of the Illumina Linkage SNP panel³³ containing 5850 SNP markers across the genome. The results are striking in that we identified two entirely new regions of linkage with $\text{LOD} > 3.0$ (chromosomes 2q33 and 11p12), of which the chromosome 2q linkage approaches genome wide criteria for 'significant' linkage. We also observed 'suggestive' evidence for linkage ($\text{LOD} > 2.2$ in a sib-pair study) in several additional genomic regions (chromosomes 4q25, 5p12 and 10q21). Some support for linkage was also provided for several previously identified regions with linkage evidence (18q21, 20p13 and 1q41–42). These results emphasize the utility of using well characterized SNP marker sets for carrying out linkage studies, and confirm that previously used standard microsatellite panels with 10 cM intervals may be insufficiently informative for linkage studies of complex diseases, especially when parental DNA samples are not available for study.^{34,35} These results are likely to catalyze renewed interest in association mapping of disease genes in the identified regions for RA based on linkage data.

Results

Analysis of single-nucleotide polymorphism linkage scan

A total of 642 Caucasian families containing 1371 affected siblings with RA were used for the SNP linkage analysis. Table 1 describes the distribution of sibships according to the number of affected siblings and additional relatives who were genotyped. The majority of families comprise a single sib pair for whom parents could not be recruited, and no other affected relative was available. A full characterization of affected individuals and clinical parameters has been reported in 512 families from this collection.³⁶ The median age at onset for the affected population studied in the current report was 39 years, with an interquartile range of 30–49 years. The mean duration of disease in this population is 16 years with a median of 13 years. We also analyzed the data according to demographic and clinical subgroups. For this analysis, we divided the families into all possible pairs of affected individuals (including primarily sib pairs but a few half-sib and other pairings as depicted in Table 1). There were

Table 1 Structure and sampling of 642 Caucasian sibling pair families studied

Number of affected sibs	Total number of families ^a	Number with 0 parents genotyped	Number with only 1 parent genotyped	Number with 2 parents genotyped
2	576 ^b	341	172	63
3	59	32	22	5
4	9	6	3	0
6	1	1	0	0

^aThree families contain sib trios in one nuclear family and a sibling pair in another nuclear family from a different generation (uncle/aunt's family) and are counted separately in each category.

^bIncludes 23 distinct half-sibship families.

55 male–male affected pairs, 284 male–female pairs and 459 female–female pairs. Similarly, among the pairs with data available on anti-CCP antibodies, there were 64 pairs concordant negative for anti CCP antibody, 230 pairs discordant for anti-CCP antibody and 438 pairs concordant positive for anti-CCP antibody, when the cut point of 20 was used to define positive and negative reactions. Furthermore, there were 59 pairs concordant negative, 112 pairs discordant and 618 pairs concordant positive for HLA SE status.

Linkage was examined using the genotypes obtained from the Illumina SNP linkage IV panel (see Methods). Table 2 summarizes all regional maximum LOD scores for each chromosome, including the pseudoautosomal regions (19 on XYp and 7 on XYq). Supplementary Figures 2–9 provide graphs of linkage analyses of chromosomes both with and without correction for having $\text{LOD} > 1.5$ linkage disequilibrium (LD). As can be seen in the table, peak LOD scores decreased on all chromosomes when markers that were in LD with each other were eliminated from analysis. When a criterion that markers showing a D' value greater than 0.7 were to be eliminated, 866 out of 5726 markers were dropped (15.12%); when only autosomal chromosomes are considered, 784 out of 5407 markers tested were dropped (14.50%). Over all markers, there was an average decrease in LOD score of 0.129 when markers in LD were dropped. For only the autosomal chromosomes, dropping markers in LD lead to a decrease of 0.120 in LOD score. Dramatic decreases in LOD scores were noted on a few chromosomes when markers in LD were dropped, for example on chromosome 21, the LOD score decreased from 11.59 to 1.11. The most prominent of these instances involved markers located at the telomeres, and can be explained by the lack of any flanking markers that are not in LD with the set of markers. Substantial reductions in LOD score also occurred in nontelomeric regions where LOD scores were greater than one; however, in these regions, significant evidence for linkage was generally maintained after dropping markers in LD. To eliminate the potential for false positive results when markers in LD are analyzed as if they are not in LD, we restricted the remainder of the analyses to only that set of markers that showed D' values less than 0.7. Using a cut point of $R^2 < 0.16$ led to fewer markers being dropped (results not shown). Even restricting markers to those with $R^2 < 0.05$ also did not qualitatively change any results. In the latter case, the

Table 2 Maximal LOD scores by chromosome and positions with (w-LD) and without (wo-LD) eliminating markers in linkage disequilibrium, by chromosome

Chromosome	LOD scores		Peak SNP – with LD		Peak SNP – without LD	
	w-LD	wo-LD	SNP	Position	SNP	Position
1	1.65	1.41	rs335523	211882862	rs1547502	216620855
2	4.02	3.52	rs1354905	189649014	rs1949429	192602000
3	0.82	0.57	rs11235	52702435	rs7061	50073527
4	3.11	2.35	rs223383	104209760	rs1384401	105023898
5	3.06	2.55	rs1857844	24569227	rs903391	43206838
6	18.53	16.14	rs169679	28964566	rs11908	32991663
7	2.49	1.94	rs903898	146244418	rs2040587	114230026
8	1.61	0.91	rs1735173	146076058	rs1375956	106521094
9	0.75	0.63	rs1138374	37964743	rs1414944	66913499
10	3.76	2.54	rs579142	60472666	rs1227938	70502871
11	3.92	3.09	rs2035693	39201341	rs1462224	41024049
12	5.09	1.46	rs7960480	132017084	rs2009625	22216512
13	0.93	0.73	rs20411	34029488	rs20411	34029488
14	0.72	0.5	rs1863766	79335457	rs1530771	79015563
15	1.14	0.33	rs1864299	47323908	rs1030588	44442237
16	1.73	1.69	rs1946155	53412897	rs1946155	53412897
17	1.74	0.97	rs411602	69136700	rs764426	67565205
18	1.9	1.47	rs1792723	52015361	rs663220	52349253
19	4.52	0.21	rs1465789	63637868	rs306450	61192862
20	3.11	1.02	rs1434789	132900	rs914433	55196924
21	11.59	1.11	rs2835629	37442056	rs2837710	40861199
22	0.34	0.32	rs139062	39124832	rs139062	39124832
X	3.22	1.17	rs2015312	56385990	rs2057652	10529107
XY ^a	8.69	3.91	rs2535444	2282688	rs700447	153441492

^aNo positive LOD scores were obtained after balancing sex concordant and discordant pairs. Abbreviations: LD, linkage disequilibrium; LOD, logarithm of odds.

LOD scores on chromosomes 2, 4, 7, 10 and 11 increased slightly with the maximum increase being from 2.35 to 2.55 on chromosome 4, while on chromosomes 5, 6, 12, 16 and 18 there were modest decreases in LOD score, with the largest decreases being on chromosome 18 from 1.47 to 1.08 and chromosome 5 from 2.55 to 2.32. The overall moderate changes in LOD scores argue strongly against false positive results due to LD after adjustment for LD ($D' \leq 0.7$) between markers. In addition, we checked the LOD scores among families with no parents genotyped versus families with at least one genotyped parent for the major regions of linkage on chromosomes 2, 4, 7, 10 and 11. Logarithm of odds scores remained positive in all family groups. On chromosomes 2, 7 and 11 the LOD scores from the families with one or more parents typed were higher, while on chromosomes 4 and 10, the LOD scores were higher for the set of families without typed parents. The only potential region of concern occurs on chromosome 4, for which the maximum LOD score in families without typed parents was 2.50, while among the families with at least one genotyped parent the LOD score was only 0.08. However, the finding that LOD scores on chromosome 4 do not decrease when restricting to R^2 values for LD among adjacent markers of 0.05 or less indicates that the evidence for linkage on this chromosome does not reflect a false positive due to LD among tightly linked markers.

Compelling new evidence for linkage on chromosomes 2q33 and 11p12

LOD scores greater than 1.5 after removing markers in LD were noted for chromosomes 2, 4, 5, 6, 7, 10, 11 and

16. Graphs of the linkage signals across these chromosomes are shown in Figure 1. Across the genome, by far the strongest evidence for linkage is found on chromosome 6p21 in the vicinity of the MHC with a maximum LOD score of 16. This is consistent with previous reports^{7,8,28–30} although the area supporting evidence for linkage is very broad, suggesting the influence of multiple susceptibility loci in this region. More strikingly, prominent and previously unreported linkage peaks are observed on chromosomes 2q33 and 11p12 with LOD scores of 3.52 and 3.09, respectively, after adjustment for LD ($D' \leq 0.7$) between markers. Furthermore, linkage evidence meeting accepted criteria for 'suggestive' linkage are also observed on chromosomes 5 and 10. Somewhat lower linkage peaks are observed on chromosomes 4, 7 and 16. Not shown in the figure are the linkage signals on chromosomes 1, 12 and 18, all of which have maximum linkage peaks that fall just below 1.5, as indicated in Table 2.

Overall, these data stand in rather dramatic contrast to what has been observed in previous microsatellite linkage scans, which have generally produced much weaker evidence for linkage outside of the MHC.^{7,8,28–31} A reason for this discrepancy in linkage results is due to the difference in the informativeness of the two markers sets. Using an entropy-based measure,³⁷ we have calculated overall informativeness at the analyzed markers for the autosomes of our previous microsatellite markers compared with the current SNP linkage set. Overall, there is a substantial increase in the information content of the SNP linkage marker set. The mean information for the Marshfield marker set 8A was 52.6%

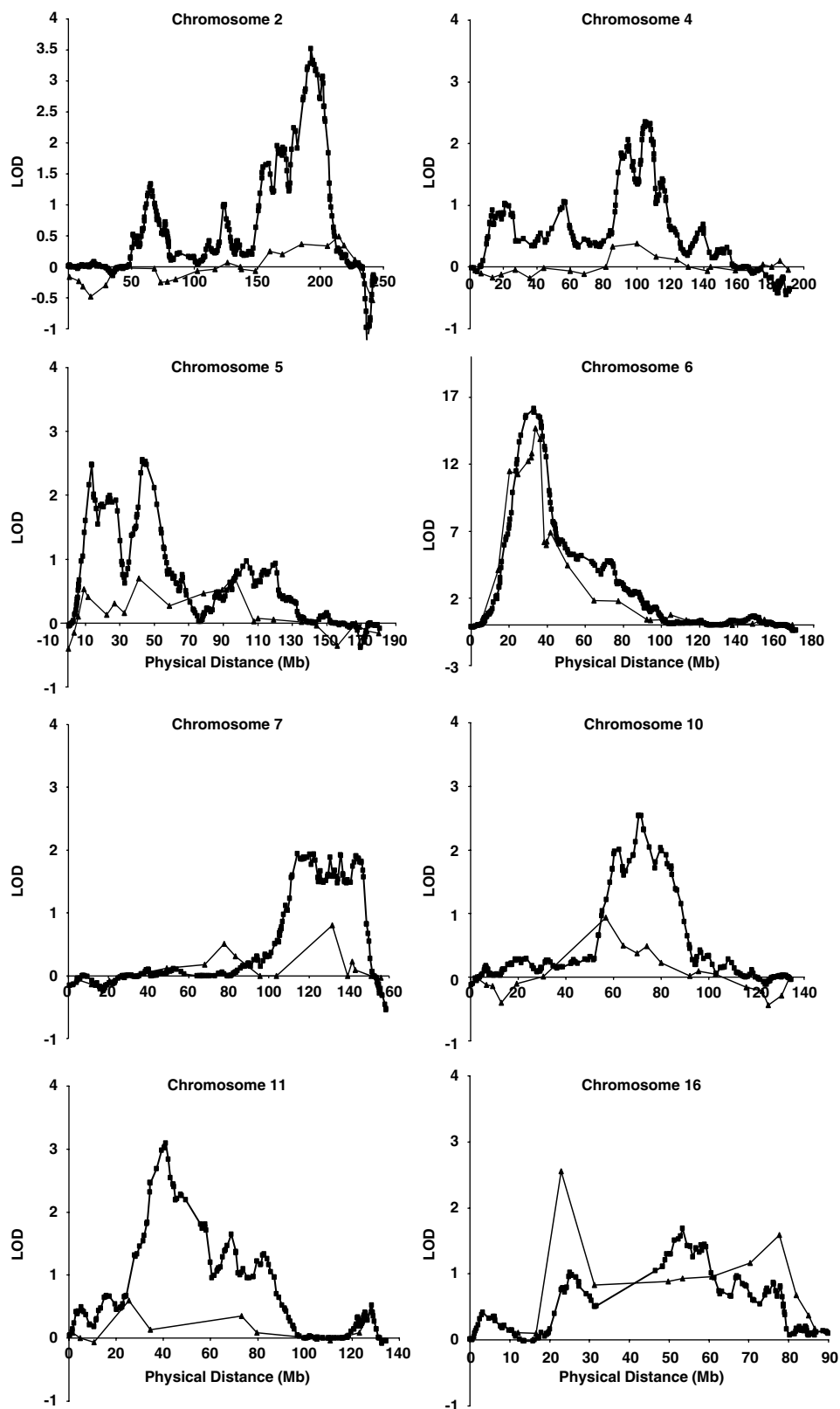


Figure 1 LOD score analysis of eight chromosomes using microsatellite and SNP panels. LOD scores are plotted for the eight chromosomes that contained maximum linkage peaks with LOD scores exceeding 1.5 on the SNP linkage scan (dark line). For comparison, the results of a previous linkage scan with microsatellites markers on 512 families⁷ is shown (light line). Note that the scale of the *y*-axis differs for chromosome 6.

(range = (0.18, 0.71), interquartile range = (0.49, 0.56) This compares with 75.6% for the current Illumina SNP set including all markers or 74.9% after excluding markers in LD with other markers (range = (0.62, 0.82) interquartile range = (0.73, 0.77)). Thus, as illustrated in Figure 1, previous linkage evidence using microsatellite markers on 512 affected sibling pair families generally showed weak evidence of linkage on all chromosomes except in the MHC on chromosome 6. This difference was observed even taking into account the somewhat smaller sample size of the previous dataset. For example, a direct comparison of the linkage results using microsatellites versus SNPs restricted to the 467 Caucasian families from the original 512 families yields LOD score increase of 2.71 for chromosome 2q33 (from a LOD score of 0.31 at 185 Mb on the microsatellite scan to 3.02 at 193 Mb on the SNP scan) and a LOD score increase of 2.04 for chromosome 11p12 (from a LOD score of 0.21 at 35 Mb on the microsatellite scan to 2.25 at 41 Mb on the SNP scan). On chromosome 6, where we had a denser microsatellite map,⁷ the information content was 69%, and there was a relatively smaller change percent change in LOD score from 14.66 with microsatellites to 16.14 with SNPs (after excluding those in LD with $D' > 0.7$).

Influence of anti-CCP antibody phenotype on linkage evidence

It has recently become apparent that the anti-CCP antibodies are a preclinical risk factor for the future development of RA, and that these antibodies are quite specific for the disease.³⁸ There is a strong association with the presence of anti-CCP antibodies and the HLA 'SE' alleles that are classically associated with RA.^{39,40} Furthermore, anti-CCP antibodies are predictive of progressive erosive disease.⁴¹ Thus, these antibodies may mark a genetically more homogeneous subset of RA. We therefore examined the effect of this trait on linkage evidence. As summarized in Table 3, evidence of significant linkage variability for the primary linkage peak was present for chromosome 6, consistent with previous reports. In addition, an increased linkage evidence is observed in siblings concordant for elevated anti-CCP titers on chromosomes 4, 5 and 7. Thus, it appears likely that follow-up association studies on the anti-CCP+ disease subset would be a rational strategy for fine mapping and gene identification.

Table 3 Tests for significant heterogeneity according to anti-CCP antibody status

Chromosome	Mean LOD scores per pair for anti-CCP strata (number of pairs)			P-value
	Negative (n = 64)	Discordant (n = 230)	Concordant (n = 438)	
2	-0.0067	0.0200	0.0262	0.446
4	-0.0305	0.0187	0.0334	0.029
5	-0.0402	0.0379	0.0188	0.032
6	0.0884	0.0270	0.0792	0.003
7	-0.0400	0.0203	0.0376	0.030
10	0.0076	0.0088	0.0286	0.275
11	0.0124	0.0290	0.0282	0.904
12	0.0207	0.0113	0.0190	0.818
16	0.0490	0.0293	-0.0000	0.133
18	0.0120	0.0389	0.0022	0.208

Analysis of sex and human lymphocyte antigen type as covariates for linkage

We also examined evidence for interactions between linkage signals and sex and/or HLA SE status. Not surprisingly, HLA SE status had a major impact on linkage evidence ($P < 0.0001$) within the MHC. The mean LOD scores per sib pair were 0.066 among pairs concordant for nonsharing of SE, -0.118 among pairs discordant for SE and 0.093 among pairs concordant for SE. The total LOD scores across strata are 3.91 among 59 pairs that are SE concordant negative, -13.23 among 112 pairs discordant for SE and 57.20 among 618 pairs concordant for SE. Thus, the SE does not explain all of the linkage within this region, as we have recently reported using data from microsatellites.³⁹ No other chromosomes exhibited significant ($P < 0.05$) heterogeneity when HLA SE status was considered. Consideration of sex as a covariate did not significantly alter linkage evidence, except at the MHC ($P = 0.0015$). Male-male pairs showed the strongest linkage signal on chromosome 6p with a mean LOD score of 0.142, male-female pairs had a mean LOD of 0.076 and female-female pairs had a mean LOD score of 0.049. The total LOD scores across strata are 7.82 among 55 male-male pairs, 21.48 among 284 pairs male-female pairs and 22.46 among 459 pairs concordant female pairs. These results, along with those from studying anti-CCP concordant and discordant pairs show a striking influence of certain covariates on linkage evidence particularly in the HLA region.

Discussion

In this study we have undertaken a dense SNP linkage scan on the largest collection of RA affected sibling pair families reported to date. The results are remarkable in that, compared with previous linkage studies using microsatellites, dramatic new evidence for several regions of linkage has emerged.

The application of linkage approaches to mapping genes for complex disorders has met with mixed success over the last decade. The identification of NOD2/CARD15 as a susceptibility gene for Crohn's Disease^{42,43} is perhaps the best example of using linkage for positional mapping of an important susceptibility gene for a complex autoimmune disorder, but it has been unclear whether this simply reflected the genetic equivalent of 'low hanging fruit'. With some exceptions, the frequent lack of replication of linkage signals in RA has dampened enthusiasm for this approach, and a recent dense linkage SNP linkage scan using 10 548 SNPs using the Affymetrix platform did not substantially change the situation,²⁹ perhaps reflecting the limited sample size that was available in that study.

While candidate gene association studies have yielded somewhat better results than linkage for identifying susceptibility genes for RA, the type 1 error rate of published association studies is high.⁴⁴ Outside of the MHC, the PTPN22 620W allele is the most compelling replicated genetic association with RA.²⁷ The discovery of the PTPN22 association with RA was based in part on positional linkage information,^{13,18} but was also informed by a broad candidate gene approach. A candidate gene association approach was also used to identify the PTPN22 association with type 1 diabetes¹⁹

and subsequently other related autoimmune disorders.^{21,45} Although the utility of linkage for disease mapping in complex autoimmune disorders remains uncertain, the data reported here are likely to reinvigorate efforts to map susceptibility genes for RA based on positional information.

Our analysis indicates that the improved linkage signals we observed using the Illumina SNP marker set resulted from the fact that the information content across the entire genome was 44% higher using the SNP panel as opposed to the microsatellite marker set from Marshfield.⁷ This increase in marker informativeness is especially important in datasets where parental information is missing,³⁴ as in late onset diseases like RA. As shown in Table 1, parental DNA was available on only 41.1% of our sibling pair families, and only 10.5% of families had both parents available for genotyping. As shown in Figure 1, for all regions showing linkage evidence with LOD > 1.5 in these families using SNP markers, the previous microsatellite analysis revealed weak evidence of linkage. This can be explained by the somewhat smaller number of families ($n=512$) previously analyzed,⁷ a possible influence from ethnic heterogeneity in our previous studies that included nonCaucasians, as well as the higher information content of the current study.

Rheumatoid arthritis is a clinically heterogeneous disorder with variability in severity, disease course and response to therapy. The recent identification of anti-CCP antibodies has provided an important tool for subdividing the phenotype.³⁸ Rheumatoid arthritis patients with these antibodies tend to have more severe progressive disease.⁴¹ In addition, anti-CCP antibodies are strongly associated with HLA alleles, both positively with SE alleles^{39,46} and negatively with HLA-DR3.^{40,47} Thus, it is not surprising that linkage evidence at the MHC is strongly influenced by the presence of these autoantibodies.³⁹ Interestingly, our data provide some evidence that other regions, particularly on chromosomes 4, 5 and 7, may also carry susceptibility alleles that are more closely linked to anti-CCP + disease. In contrast, neither sex nor HLA status significantly influenced linkage results outside of the MHC.

The linkage results at the MHC are remarkable not only for their strength, but also for the very broad linkage peak. Not all of this linkage can be ascribed to the known HLA associations with 'SE' alleles encoded at the HLA-DRB1 locus. Indeed, there is previous evidence that genes within the central MHC also contribute to disease risk, independent of DRB1.^{10,11,48} The current data suggest that additional risk genes may also lie within the very broad linkage peak in this region, perhaps some of which lie outside of the MHC proper.

The strong linkage signal on chromosome 11p is unexpected. Interestingly, the peak of the linkage signal at 41 Mb lies within a 'gene desert' in which only one known functional gene is present in a region of over 6 Mb (~36.6–43.0 Mb, build 35). The sole gene in this region is NGL-1, a ligand for netrin G1, a molecule that is involved in axon guidance in the developing central nervous system. NGL-1 is predominantly expressed in brain tissue, but is also expressed at lower levels in a variety of other tissues, including peripheral blood.⁴⁹ Interestingly, the RAG1, RAG2 and TRAF6 genes are located immediately telomeric

to this large gene desert, well within the 1-LOD support interval.

The linkage peak on chromosome 2q33 is located in the region containing the ICOS-CD28-CTLA4 gene cluster. Previous association studies have shown some evidence of association CTLA4 with RA,⁵⁰ and we have recently confirmed a weak effect at this locus using several large case-control data sets.²⁷ Interestingly, the association with the CTLA4 CT60 polymorphism is stronger in the anti-CCP + disease subset.²⁷ However, the strength of the CTLA4 C60 association with RA is not adequate to explain the entire linkage signal reported here, and thus we anticipate that additional risk alleles, perhaps in different genes, will be found in this region.

Comparisons of our findings with previous genome-wide linkage analyses indicate some regions of overlap, outside the MHC. Osorio *et al.*³⁰ performed a dense microsatellite linkage analysis of 88 French Caucasians families that included 105 sibling pairs. Regions of overlap from that study with those we have identified as yielding LOD scores over 1.0 include 1q44, 2q33, 18q21, 20p13 and 12p (our peak linkage region is at 12p12 about 10 cM from their peak region). John *et al.*²⁹ performed a dense SNP scan of 157 multicase RA families and identified several regions showing evidence for linkage at the $P < 0.05$ level, but our study yielded only weak evidence for linkage to those regions except the 6p, 6q and 16p12 (LOD = 1.0 at rs874562) regions. A genome-wide linkage analysis of 41 Japanese families⁵¹ identified significant linkages to chromosomes 1p36, 8q22–23 and Xq27. Of these loci, we did not detect evidence for linkage to 1p36 or Xq27 but our peak evidence (LOD = 0.91) on chromosome 8 occurs at rs1375956, which is in the 8q22–23 region. Meta-analyses³² and joint analyses⁵² of the different data sets will help to resolve similarities and differences among the different populations. Studies on US populations include a heterogeneous collection of ethnic groups. We have attempted to minimize this heterogeneity by restricting our study to include only Caucasians. However, even among European groups, considerable heterogeneity with respect to causal factors such as PTPN22 may exist. Therefore, future studies that condition on subethnic origin may help to identify any specific factors that influence risk differentially among different subethnic groups.

Analysis of linkage to specific candidate loci showed weak evidence for linkage for most loci outside of the MHC (Supplementary Figures 2–9). Among all candidate loci on chromosomes with LOD scores over 1.0 (PADI4, CTLA4, PTPN22, RUNX1, SLCA224, HLADRB1, IL2, SLC11A1, TNFRSF1A, IL26, MHC2TA and FCR Gamma 3a), model free LOD scores over 0.5 were only obtained for HLA-DRB1 (LOD = 16.1), CTLA4 (LOD = 2.35), IL2 (LOD = 0.52) and RUNX1 (LOD = 0.67). However, given the weak impact that the currently known non-MHC candidates have upon risk for RA, failure to detect linkages is not surprising. The most consistently replicated non-MHC locus is PTPN22, which was reported¹³ to confer an increased risk of 1.97 per risk allele with an allele frequency of 0.09 in healthy Caucasians. Applying these values (assuming a log-additive model for risk to homozygous carriers) and assuming an overall lifetime prevalence of 0.7%, we have genotype specific penetrances of 0.006, 0.012 and 0.023 for normal,

heterozygous and homozygous at risk genotypes. With these parametric values, we calculated the expected LOD using the positions, allele frequencies and the missing data patterns four markers immediately adjacent to the PTPN22. The median expected LOD score was average expected LOD score using these markers is 1.70 ± 1.13 with a range of -0.17 to 7.25 with 30% having LOD scores of 1.0 or less and 66% of replicates having LOD scores of 2 or less. Thus, our failure to identify the PTPN22 locus appears to reflect limited power to detect a locus having the limited relative risk associated with the R620W allele. In general, we anticipate that loci that are positionally identified through linkage analysis would confer higher risks than the approximately two-fold increased risk provided by the R620W allele of PTPN22.

Overall, our results provide compelling support for the view that multiple risk genes with modest effect contribute to risk for RA, and that these genes are amenable to identification by combining both linkage and association approaches to positional mapping, supplemented by the selection of candidate genes based on the evolving knowledge of the pathophysiology of autoimmunity as well as expression array studies.^{49,53} The data illustrate the striking increase in linkage information that can be obtained by using appropriate SNP linkage panels. Recently, a number of reports have appeared using such panels, with evident success in identifying new regions of linkage. Intriguingly, a recent SNP linkage analysis of familial chronic lymphocytic leukemia (CLL) provides evidence for linkage at 11p12, overlapping with the linkage peak reported here for RA.⁵⁴ Perhaps this indicates the involvement of common B cell genes in these disorders. Notably, CLL is often accompanied by the presence of autoantibodies, and current views of B-CLL pathogenesis invoke B cell autoreactivity as contributing to tumor development.⁵⁵ In contrast, an SNP linkage scan of multiple sclerosis has not revealed dramatic new evidence for linkage,⁵⁶ perhaps suggesting an even more complex and/or less strong genetic component to this disorder. Other major human autoimmune diseases with a clear genetic component, such as type 1 diabetes and systemic lupus erythematosus have not yet been studied using linkage panels based on SNP markers. It will be of great interest to see how such studies will compare with the results reported here, since these two disorders share some degree of overlapping genetic susceptibility with RA.

Methods

Study population

Details of enrollment procedures for NARAC have been published previously.⁷ The detailed clinical and marker data and for the first two microsatellite screens on 512 families are available on the NARAC website (<http://www.naracdata.org>). The clinical data for the additional sibling pair families reported here are also given on the NARAC website. The geographic and clinical information, including ethnic background, age at onset of RA symptoms was collected from all NARAC RA patients by telephone. Confirmation of RA diagnosis, including which ACR criteria were fulfilled, was obtained from patients' rheumatologists. Radiographs of the hands and wrists were also obtained to document the presence and

extent of joint involvement. Eligible families had to meet the following criteria: (1) two or more siblings satisfied the 1987 American College of Rheumatology (ACR) criteria for RA;⁵⁷ (2) at least one sibling had documented erosions on hand radiographs; and (3) at least one sibling had disease onset between the ages of 18 and 60 years. The presence of other diseases that are accompanied by inflammatory arthritis, such as psoriasis or inflammatory bowel disease, was an exclusionary criterion for families. Informed consent was obtained from every subject, including all participating family members, and approval of the local institutional review board was secured at every recruitment site prior to enrollment. Overall, we have studied 1592 affected siblings contained within 749 multicase RA families recruited as a collaborative effort of the NARAC. Of these, 1371 affected siblings in 642 multicase Caucasian families have been retained for the current analysis. Non-Caucasian families have been excluded from the current report in order to limit genetic heterogeneity of the study population. Non-Caucasian families have been excluded from the current report in order to limit genetic heterogeneity of the study population. Two Caucasian families – families 03006 and 07022 – were included as RA affected sibling pairs, even though criteria for erosive disease were not met; in each of these cases, at least one sibling was CCP+. Family 05006 was analyzed as having three affected siblings but individual 202 did not have documented RA.

Anti-CCP testing

NARAC serum samples collected at study entry were tested for anti-CCP autoantibodies at the University of Washington, Department of Laboratory Medicine, Immunology Division. Anti-CCP titers were determined based on a second-generation enzyme-linked immunosorbent assay (ELISA) (INOVA Diagnostics). Anti-CCP levels exceeding 20 were considered to show a positive antibody titer while values less than this were categorized as negative.

Genotyping using the Illumina linkage IV single-nucleotide polymorphism set

We utilized a set of 5858 SNPs distributed across the genome for our linkage panel (for details, see www.illumina.com). This panel contains 5858 SNP makers. Of these, 98.1% passed quality control filters, yielding a total data set of 5744 SNPs. An additional 18 markers on the Y chromosome were dropped as uninformative for linkage analysis. Overall, on the basis of Mendelian inconsistencies, sporadic error rates were estimated at 0.0008%. Of the 642 Caucasian families, 468 had previously been typed with microsatellite markers⁷ so that errors due to sample handling and/or non-paternity had been previously eliminated. Among the 174 newly typed families, 15 cases of definite or presumed (half sibs when parents not available) non-paternity were detected. Among the 4995 markers on the autosomes and excluding chromosome 6 (because of strong genetic factors on it), only five markers had evidence of being out of Hardy–Weinberg equilibrium with $P \leq 0.001$ when using parental genotypes, which conforms well to the expected frequency of Hardy–Weinberg disequilibrium with this large set of markers. Of note, marker rs238510 at 103.49 Mb in the linkage peak on chromosome 4 and just proximal to the potential candidate gene, B-cell scaffold protein with

ankyryn repeats 1 (BANK1), showed a significant ($P \leq 0.001$) departure from Hardy–Weinberg equilibrium. Since genetic maps are not available for the majority of SNPs that were available in this panel, we assumed that 1 Mb is 1 cM. Ulgen and Li⁵⁸ have found in simulated data that for dense mapping scenarios, applying the 1 cM equals 1 Mb rule versus typing the map distances to known anchor positions, did not substantively alter the evidence for linkage. In the proximity of the centromeric regions, this mapping approach is inaccurate because the centromeric regions have suppressed recombination. Therefore, we reperformed analysis setting the recombination to zero for the six chromosomes with large centromeres (1, 3, 9, 11, 16, 19). Excluding recombination in these regions led to slightly lower LOD scores (further results available upon request).

Human lymphocyte antigen-DRB1 genotyping

Broad-level HLA-DRB1 typing for the allelic groups DRB1*01 through DRB1*18, and high-resolution DRB1*04 typing were accomplished by initial PCR amplification of groups of alleles using biotinylated PCR primers, followed by hybridization to immobilized sequence-specific oligonucleotide probes in a linear array format. Positive hybridization reactions were detected using a streptavidin-horseradish peroxidase conjugate and a soluble colorless substrate TMB (3,3',5,5'-tetramethylbenzidine).⁵⁹ A computer algorithm based on the sequence specific oligonucleotide probe hybridization pattern and the Anthony Nolan 1999 HLA sequence database (<http://www.ebi.ac.uk/imgt/hla/>) was used to assign genotypes. The following alleles detected were classified as 'SE' positive: DRB1*0101, 0102, 0104, 0105, 0401, 0404, 0405, 0408, 0409, 1001, 1402 and 1406.

Statistical methods

Allele frequencies were estimated from the direct observations of alleles at each locus. The very large number of individuals in our study ensured that the allele frequencies so obtained would be sufficiently precise, even though we did not adjust for familial relationships in the frequency estimation process. We used the exact test for Hardy–Weinberg equilibrium implemented in PEDSTATS to identify loci that showed strong disequilibrium ($P \leq 0.001$). In regions showing evidence for linkage (LOD > 1.5), linkage analyses were performed including and excluding markers that strongly deviated from Hardy–Weinberg equilibrium. Departures from Hardy–Weinberg equilibrium can occur for numerous reasons including genotype call failures and association between marker alleles and disease susceptibility. Since the latter would be expected for certain regions showing strong evidence for linkage, we here only report results including all markers.

Linkage analysis of tightly linked loci can lead to an excess of false positive results if the markers are in strong LD and parents are not available for genotyping.⁶⁰ Therefore, we applied the suite of Perl scripts implemented in SNPLINK⁶¹ that has been developed and implemented for processing high-throughput linkage analyses, while allowing the user to remove markers that are jointly in LD. We choose to remove markers that showed (Lewontin) LD D' values greater than 0.7, because our earlier analyses on simulated data have shown little inflation in LOD score if this criterion is

used.⁶⁰ The SNPLINK software computes, internally, measures of LD for pairs of markers. Haplotype blocks are then formed according to the method of Gabriel *et al.*⁶² The software then removes markers that are in LD with the central marker within any haplotype block showing LD. The SNPLINK software permits the user to choose among software programs for analysis, and we used Merlin⁶³ to compute nonparametric Kong and Cox LOD scores.⁶⁴ These LOD scores summarize the evidence in favor of linkage, expressed as the \log_{10} likelihood ratio of the data assuming linkage versus the model assuming no linkage. The testing procedure weights data from each position according to the informativeness of the marker data at the position. We used a model-free test for linkage for these analyses because the etiology of RA appears to be complex. For analysis using Merlin, we used the all pairs option, which is a model free linkage test with optimal power to detect dominantly acting loci.⁶⁵ We also removed data from any families for which a preliminary analysis suggested the strong possibility of double recombinants as discussed in the Merlin documentation (<http://www.sph.umich.edu/csg/abecasis/Merlin/tour/error.html>). Following this procedure, 2469 autosomal SNP genotypes were identified as possibly erroneous, which leads to an estimated unlikely genotype rate of $2469 / (642 \times 2.78 \times 5414) = 0.026\%$, since 2.78 individuals were genotyped per family. We also evaluated the positions of these possible errors to identify any regions for which the SNP map could have been erroneous (for example), which could cause multiple errors in the same region. We did not find any such regions and also failed to find any markers showing a particularly high number of possible double recombinants.

We also reanalyzed the data using a slightly more restrictive criterion that the R^2 value between markers was less than 0.05. Because results were very similar to those we found when eliminating markers with $D' > 0.7$, we did not present them.

Linkage analysis of the pseudoautosomal regions poses special issues when there is a skewing of the sex ratio among the individuals being studied. Selection of sib pairs that include an excess of same-sex pairs leads to an excess of sharing in the pseudoautosomal region. Among both male–male and female–female pairs, there is on average an excess of identity by descent sharing within this region, while opposite sex-pairs show a decrease in sharing. Selection for either sex therefore increases the identity by descent sharing in the region. To allow for this skewing and achieve an equal number of sex concordant and sex discordant pairs, we performed one analysis eliminating male–male pairs as well as 37% of female–female pairs, and we performed a second analysis including male–male pairs but eliminating 45% of female pairs.

To evaluate the hypothesis that linkage evidence may vary among strata defined by elevated anti-CCP titers, sex, or SE, we performed a nonparametric, Kruskal–Wallis analysis of variance treating the LOD scores as the dependent observations and the strata as independent predictors.⁶¹

Acknowledgements

This work was supported by the National Institutes of Health (RO1-AR44222 and Contract NO1-AR-2-2263)

and by the National Arthritis Foundation. We thank Joseph Garsetti and Mark Hansen at Illumina Inc. for their efforts in support of this project.

References

- 1 Firestein GS. Evolving concepts of rheumatoid arthritis. *Nature* 2003; **423**: 356–361.
- 2 Seldin MF, Amos CI, Ward R, Gregersen PK. The genetics revolution and the assault on rheumatoid arthritis. *Arthritis Rheum* 1999; **42**: 1071–1079.
- 3 MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K *et al*. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum* 2000; **43**: 30–37.
- 4 Risch N. Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* 1990; **46**: 222–228.
- 5 Gregersen PK. Genetics of rheumatoid arthritis: confronting complexity. *Arthritis Res* 1999; **1**: 37–44.
- 6 Hall FC, Hall FC, Weeks DE, Camilleri JP, Williams LA, Amos N *et al*. Influence of the HLA-DRB1 locus on susceptibility and severity in rheumatoid arthritis. *Qjm* 1996; **89**: 821–829.
- 7 Jawaheer D, Seldin MF, Amos CI, Chen WV, Shigeta R, Etzel C, *et al*. North American Rheumatoid Arthritis Consortium. Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families. *Arthritis Rheum* 2003; **48**: 906–916.
- 8 Cornelis F, Faure S, Martinez M, Prud'homme JF, Fritz P, Dib C, *et al*. ECRAF. New susceptibility locus for rheumatoid arthritis suggested by a genome-wide linkage study. *Proc Natl Acad Sci USA* 1998; **95**: 10746–10750.
- 9 Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum* 1987; **30**: 1205–1213.
- 10 Mulcahy B, Waldron-Lynch F, McDermott MF, Adams C, Amos CI, Zhu DK *et al*. Genetic variability in the tumor necrosis factor-lymphotoxin region influences susceptibility to rheumatoid arthritis. *Am J Hum Genet* 1996; **59**: 676–683.
- 11 Jawaheer D, Li W, Graham RR, Chen W, Damle A, Xiao X *et al*. Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. *Am J Hum Genet* 2002; **71**: 585–594.
- 12 du Montcel ST, Michou L, Petit-Teixeira E, Osorio J, Lemaire I, Lasbleiz S *et al*. New classification of HLA-DRB1 alleles supports the shared epitope hypothesis of rheumatoid arthritis susceptibility. *Arthritis Rheum* 2005; **52**: 1063–1068.
- 13 Begovich AB, Carlton VE, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC *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* 2004; **75**: 330–337.
- 14 Lee AT, Li W, Liew A, Bombardier C, Weisman M, Massarotti EM *et al*. The PTPN22 R620W polymorphism associates with RF positive rheumatoid arthritis in a dose-dependent manner but not with HLA-SE status. *Genes Immun* 2005; **6**: 129–133.
- 15 Simkins HM, Merriman ME, Highton J, Chapman PT, O'Donnell JL, Jones PB *et al*. Association of the PTPN22 locus with rheumatoid arthritis in a New Zealand Caucasian cohort. *Arthritis Rheum* 2005; **52**: 2222–2225.
- 16 van Oene M, Wintle RF, Liu X, Yazdanpanah M, Gu X, Newman B *et al*. Association of the lymphoid tyrosine phosphatase R620W variant with rheumatoid arthritis, but not Crohn's disease, in Canadian populations. *Arthritis Rheum* 2005; **52**: 1993–1998.
- 17 Seldin MF, Shigeta R, Laiho K, Li H, Saila H, Savolainen A *et al*. Finnish case-control and family studies support PTPN22 R620W polymorphism as a risk factor in rheumatoid arthritis, but suggest only minimal or no effect in juvenile idiopathic arthritis. *Genes Immun* 2005; **6**: 720–722.
- 18 Gregersen PK. Pathways to gene identification in rheumatoid arthritis: PTPN22 and beyond. *Immunol Rev* 2005; **204**: 74–86.
- 19 Bottini N, Musumeci L, Alonso A, Rahmouni S, Nika K, Rostamkhani M *et al*. A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat Genet* 2004; **36**: 337–338.
- 20 Onengut-Gumusc S, Ewens KG, Spielman RS, Concannon P. A functional polymorphism (1858C/T) in the PTPN22 gene is linked and associated with type I diabetes in multiplex families. *Genes Immun* 2004; **5**: 678–680.
- 21 Velaga MR, Wilson V, Jennings CE, Owen CJ, Herington S, Donaldson PT *et al*. The codon 620 tryptophan allele of the lymphoid tyrosine phosphatase (LYP) gene is a major determinant of Graves' disease. *J Clin Endocrinol Metab* 2004; **89**: 5862–5865.
- 22 Kyogoku C, Langefeld CD, Ortmann WA, Lee A, Selby S, Carlton VE *et al*. Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE. *Am J Hum Genet* 2004; **75**: 504–507.
- 23 Reddy MV, Johansson M, Sturfelt G, Jonsen A, Gunnarsson I, Svenungsson E *et al*. The R620W C/T polymorphism of the gene PTPN22 is associated with SLE independently of the association of PDCD1. *Genes Immun* 2005; **6**: 658–662.
- 24 Hinks A, Barton A, John S, Bruce I, Hawkins C, Griffiths CE *et al*. Association between the PTPN22 gene and rheumatoid arthritis and juvenile idiopathic arthritis in a UK population: further support that PTPN22 is an autoimmunity gene. *Arthritis Rheum* 2005; **52**: 1694–1699.
- 25 Viken MK, Amundsen SS, Kvien TK, Boberg KM, Gilboe IM, Lilleby V *et al*. Association analysis of the 1858C>T polymorphism in the PTPN22 gene in juvenile idiopathic arthritis and other autoimmune diseases. *Genes Immun* 2005; **6**: 271–273.
- 26 Carlton VE, Hu X, Chokkalingam AP, Schrodi SJ, Brandon R, Alexander HC *et al*. PTPN22 genetic variation: evidence for multiple variants associated with rheumatoid arthritis. *Am J Hum Genet* 2005; **77**: 567–581.
- 27 Plenge RM, Padyukov L, Remmers EF, Purcell S, Lee AT, Karlson EW *et al*. Replication of putative candidate-gene associations with rheumatoid arthritis in 14,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. *Am J Hum Genet* 2005; **77**: 1044–1060.
- 28 MacKay K, Eyre S, Myerscough A, Milicic A, Barton A, Laval S *et al*. Whole-genome linkage analysis of rheumatoid arthritis susceptibility loci in 252 affected sibling pairs in the United Kingdom. *Arthritis Rheum* 2002; **46**: 632–639.
- 29 John S, Shephard N, Liu G, Zeggini E, Cao M, Chen W *et al*. Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *Am J Hum Genet* 2004; **75**: 54–64.
- 30 Osorio Y, Fortea J, Bukulmez H, Petit-Teixeira E, Michou L, Pierlot C *et al*. Dense genome-wide linkage analysis of rheumatoid arthritis, including covariates. *Arthritis Rheum* 2004; **50**: 2757–2765.
- 31 Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995; **11**: 241–247.
- 32 Fisher SA, Lanchbury JS, Lewis CM. Meta-analysis of four rheumatoid arthritis genome-wide linkage studies: confirmation of a susceptibility locus on chromosome 16. *Arthritis Rheum* 2003; **48**: 1200–1206.
- 33 Murray SS, Oliphant A, Shen R, McBride C, Steeke RJ, Shannon SG *et al*. A highly informative SNP linkage panel for human genetic studies. *Nat Methods* 2004; **1**: 113–117.
- 34 Sawcer SJ, Maranian M, Singlehurst S, Yeo T, Compston A, Daly MJ *et al*. Enhancing linkage analysis of complex disorders: an evaluation of high-density genotyping. *Hum Mol Genet* 2004; **13**: 1943–1949.

- 35 Rijsdijk FV, Sham PC. Estimation of sib-pair IBD sharing and multipoint polymorphism information content by linear regression. *Behav Genet* 2002; **32**: 211–220.
- 36 Jawaheer D, Lum RF, Amos CI, Gregersen PK, Criswell LA. Clustering of disease features within 512 multicase rheumatoid arthritis families. *Arthritis Rheum* 2004; **50**: 736–741.
- 37 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996; **58**: 1347–1363.
- 38 Schellekens GA, Visser H, de Jong BA, van den Hoogen FH, Hazes JM, Breedveld FC *et al*. The diagnostic properties of rheumatoid arthritis antibodies recognizing a cyclic citrullinated peptide. *Arthritis Rheum* 2000; **43**: 155–163.
- 39 Huizinga TW, Amos CI, van der Helm-van Mil AH, Chen W, van Gaalen FA, Jawaheer D *et al*. Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins. *Arthritis Rheum* 2005; **52**: 3433–3438.
- 40 Irigoyen P, Lee AT, Wener MH, Li W, Kern M, Batliwalla F *et al*. Regulation of anti-cyclic citrullinated peptide antibodies in rheumatoid arthritis. *Arthritis Rheum* 2005; **52**: 3813–3818.
- 41 van Gaalen FA, Linn-Rasker SP, van Venrooij WJ, de Jong BA, Breedveld FC, Verweij CL *et al*. Autoantibodies to cyclic citrullinated peptides predict progression to rheumatoid arthritis in patients with undifferentiated arthritis: a prospective cohort study. *Arthritis Rheum* 2004; **50**: 709–715.
- 42 Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R *et al*. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 2001; **411**: 603–606.
- 43 Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J *et al*. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001; **411**: 599–603.
- 44 Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet* 2001; **29**: 306–309.
- 45 Criswell LA, Pfeiffer KA, Lum RF, Gonzales B, Novitzke J, Kern M *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* 2005; **76**: 561–571.
- 46 van Gaalen FA, van Aken J, Huizinga TW, Schreuder GM, Breedveld FC, Zanelli E *et al*. Association between HLA class II genes and autoantibodies to cyclic citrullinated peptides (CCPs) influences the severity of rheumatoid arthritis. *Arthritis Rheum* 2004; **50**: 2113–2121.
- 47 Verpoort KN, van Gaalen FA, van der Helm-van Mil AH, Schreuder GM, Breedveld FC, Huizinga TW *et al*. Association of HLA-DR3 with anti-cyclic citrullinated peptide antibody-negative rheumatoid arthritis. *Arthritis Rheum* 2005; **52**: 3058–3062.
- 48 Newton JL, Harney SM, Timms AE, Sims AM, Rockett K, Darke C *et al*. Dissection of class III major histocompatibility complex haplotypes associated with rheumatoid arthritis. *Arthritis Rheum* 2004; **50**: 2122–2129.
- 49 Batliwalla FM, Baechler EC, Xiao X, Li W, Balasubramanian S, Khalili H *et al*. Peripheral blood gene expression profiling in rheumatoid arthritis. *Genes Immun* 2005; **6**: 388–397.
- 50 Shiozawa S, Hayashi S, Tsukamoto Y, Goko H, Kawasaki H, Wada T *et al*. Identification of the gene loci that predispose to rheumatoid arthritis. *Int Immunol* 1998; **10**: 1891–1895.
- 51 John S, Shephard N, Chen S, Etzel C, Jawaheer D, Seldin M *et al*. Linkage analysis of rheumatoid arthritis in US and UK families reveals interactions between HLA-DRB1 and loci on chromosome 6q and 16p. *Arthritis Rheum* 2006, in press.
- 52 Rodriguez MR, Nunez-Roldan A, Aguilar F, Valenzuela A, Garcia A, Gonzalez-Escribano MF. Association of the CTLA4 3' untranslated region polymorphism with the susceptibility to rheumatoid arthritis. *Hum Immunol* 2002; **63**: 76–81 (2002).
- 53 Abbas AR, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F *et al*. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes Immun* 2005; **6**: 319–331.
- 54 Sellick GS, Webb EL, Allinson R, Matutes E, Dyer MJ, Jonsson V *et al*. A high-density SNP genomewide linkage scan for chronic lymphocytic leukemia-susceptibility loci. *Am J Hum Genet* 2005; **77**: 420–429.
- 55 Chiorazzi N, Rai KR, Ferrarini M. Chronic lymphocytic leukemia. *N Engl J Med* 2005; **352**: 804–815.
- 56 Sawcer S, Ban M, Maranian M, Yeo TW, Compston A, Kirby A, *et al*. International Multiple Sclerosis Genetics Consortium. A high-density screen for linkage in multiple sclerosis. *Am J Hum Genet* 2005; **77**: 454–467.
- 57 Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS *et al*. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum* 1988; **31**: 315–324.
- 58 Ulgen A, Li W. Comparing SNP-marker-based and micro-satellite-marker-based linkage analyses. *BMC Genet* 2005; **6** (Suppl 1): S13.
- 59 Erlich H, Bugawan T, Begovich AB, Scharf S, Griffith R, Saiki R *et al*. HLA-DR, DQ and DP typing using PCR amplification and immobilized probes. *Eur J Immunogenet* 1991; **18**: 33–55.
- 60 Huang Q, Shete S, Amos CI. Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *Am J Hum Genet* 2004; **75**: 1106–1112.
- 61 Webb EL, Sellick GS, Houlston RS. SNPLINK: multipoint linkage analysis of densely distributed SNP data incorporating automated linkage disequilibrium removal. *Bioinformatics* 2005; **21**: 3060–3061.
- 62 Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B *et al*. The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.
- 63 Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; **30**: 97–101.
- 64 Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet* 1997; **61**: 1179–1188.
- 65 McPeck MS. Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genet Epidemiol* 1999; **16**: 225–249.

Supplementary Information accompanies the paper on Genes and Immunity's website (<http://www.nature.com/gene>)