Published in final edited form as:

Arthritis Rheum. 2011 June ; 63(6): 1522–1526. doi:10.1002/art.30315.

Examining the Overlap Between Genome-Wide Rare Variant Association Signals and Linkage Peaks in Rheumatoid Arthritis

S. Eyre¹, X. Ke¹, R. Lawrence², J. Bowes¹, K. Panoutsopoulou³, A. Barton¹, W. Thomson¹, J. Worthington¹, and E. Zeggini²

¹S. Eyre, PhD, X. Ke, PhD, J. Bowes, PhD, A. Barton, FRCP, PhD, W. Thomson, PhD, J. Worthington, PhD: University of Manchester, Manchester, UK

²R. Lawrence, DPhil, E. Zeggini, PhD: Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK and Wellcome Trust Sanger Institute, Hinxton, UK

³K. Panoutsopoulou, PhD: Wellcome Trust Sanger Institute, Hinxton, UK

Abstract

Objective—With the exception of the major histocompatibility complex (MHC) and *STAT4*, no other rheumatoid arthritis (RA) linkage peak has been successfully fine-mapped to date. This apparent failure to identify association under peaks of linkage could be ascribed to the examination of common variation, when linkage is likely to be driven by rare variants. The purpose of this study was to investigate the overlap between genome-wide rare variant RA association signals observed in the Wellcome Trust Case Control Consortium (WTCCC) study and 11 replicating RA linkage peaks, defined as regions with evidence for linkage in >1 study.

Methods—The WTCCC data set contained 40,482 variants with minor allele frequency of ≥ 0.05 in 1,860 RA patients and 2,938 controls. Genotypes of all rare variants within a given gene region were collapsed into a single locus and a global *P* value was calculated per gene.

Results—The distribution of rare variant signals (association $P \leq 10^{-5}$) was found to differ significantly between regions with and without linkage evidence ($P = 2 \times 10^{-17}$ by Fisher's exact test). No significant difference was observed after data from the MHC region were removed or when the effect of the *HLA-DRB1* locus was accounted for.

Conclusion—The results suggest that rare variant association signals are significantly overrepresented under linkage peaks in RA, but the effect is driven by the MHC. This is the first study to examine the overlap between linkage peaks and rare variant association signals genome-wide in a complex disease.

Genome-wide association studies (GWAS) have been successfully used to identify common disease susceptibility loci for complex human diseases (1–3). In the ~30 years preceding the advent of GWAS, only variation in the major histocompatibility complex (MHC) region (4)

© 2011, American College of Rheumatology

Address correspondence to E. Zeggini, PhD, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1HH, UK. eleftheria@sanger.ac.uk.

Drs. Eyre and Ke contributed equally to this work.

AUTHOR CONTRIBUTIONS All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Zeggini had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Eyre, Ke, Barton, Worthington, Zeggini.

Acquisition of data. Ke, Thomson, Zeggini.

Analysis and interpretation of data. Eyre, Ke, Lawrence, Bowes, Panoutsopoulou, Barton, Worthington, Zeggini.

and *PTPN22* gene (5) had been confirmed as rheumatoid arthritis (RA) susceptibility loci. In contrast, at least 19 novel RA susceptibility loci have been identified over the last 2 years (6). These novel loci are represented by common (minor allele frequency [MAF] >0.05) single-nucleotide polymorphisms (SNPs) with modest or small effect sizes and together explain only a small proportion of disease variance (7). RA is believed to have a heritability of 50–60%, with the MHC region harboring approximately one-third of the genetic contribution (8). The 19 novel RA loci account for <5% of disease susceptibility variability. This leaves a large proportion of genetic variance not yet explained, which could potentially be ascribed to low-frequency and rare variation of larger effect size.

There is accumulating evidence that rare but higher-penetrance sequence variants can play an important role in the etiology of some common diseases (9,10). Indeed, large sequencing projects, such as the 1000 Genomes Project (<http://www.1000genomes.org>), are currently being conducted to discover and catalog low-frequency sequence variants. Current GWAS design and SNP content on a GWAS array do not provide sufficient power to detect disease associations at variants with low MAF. To overcome this, several methods have been proposed to cumulatively analyze rare variants (11,12). Application of such algorithms to current GWAS data sets may reveal regions harboring low-frequency or rare disease susceptibility variants.

Attempts to identify RA susceptibility loci prior to the advent of GWAS used family-based linkage studies. Despite a number of regions with replicating evidence for linkage, only 2 RA linkage peaks have been successfully fine-mapped to date. The MHC linkage peak has been ascribed to the *HLA-DRB1* locus, although additional (smaller) effects have also been localized to this region (13). More recently, common SNP associations were found in the *STAT4* locus under the linkage peak on chromosome 2q (14). No further genes or SNP variants have been mapped to other regions of linkage in RA.

The apparent failure to identify association under peaks of linkage for RA and other complex diseases may occur because currently, only common frequency variation has been examined, and linkage peak signals are likely to be caused by multiple rare variants. To address this issue, we investigated the overlap between rare variant association signals using RA genome-wide association data from the Wellcome Trust Case Control Consortium (WTCCC) study and replicating linkage peaks.

PATIENTS AND METHODS

Patients and controls

The RA patient and control study populations were as described in the initial report of the WTCCC study (1). Briefly, 2,000 RA patients were recruited for studies that had been coordinated by the Arthritis Research Campaign Epidemiology Unit. All subjects were Caucasian and over 18 years old, and satisfied the 1987 American College of Rheumatology criteria for RA (15). Of the 3,000 controls recruited for the WTCCC study, 1,500 were from the 1958 Birth Cohort (also known as the National Child Development Study), which includes all births in England, Wales, and Scotland, during 1 week in 1958. The second set of controls was made up of 1,500 individuals selected from among blood donors recruited as part of a collaboration between the WTCCC and the UK Blood Services (National Health Service Blood and Transfer in England, Scottish National Blood Transfusion Service in Scotland, and Welsh Blood Service in Wales).

Genotyping

The samples from the WTCCC were genotyped using the Affymetrix GeneChip 500K array. Sample collection and quality control procedures were performed, and poor-quality SNPs

and samples were excluded based on the criteria described in detail in the original WTCCC study report (1). Briefly, SNP exclusion thresholds were based on extreme deviation from Hardy-Weinberg equilibrium ($P < 5.7 \times 10^{-7}$), or if the study-wide missing data rate was >0.05 for SNPs with an MAF of <0.05 or >0.01 for SNPs with an MAF of <0.05 . Samples with a study-wide missing data rate of >0.03 were excluded if they were outliers based on genome-wide heterozygosity or if they were determined to be of non-Caucasian ancestry.

After quality control, the resulting data set contained samples from 1,860 RA patients and 2,938 controls, with genotype data on 40,482 variants with an MAF of >0.05 .

Statistical analysis

Replicating linkage peaks were defined as ones that contained a microsatellite marker that was significantly associated at the $P = 0.05$ level (nonparametric linkage >1.3) in at least 1 original study and 1 meta-analysis. The extent of the linkage boundaries was defined by the extreme associated microsatellite plus 10 Mb. Details on the specific regions of linkage studies are available from the corresponding author upon request.

A total of 22,344 gene regions were defined across the autosomal genome as intervals extending from 50 kb on either side of the transcriptional start and end sites of genes. At least 2 rare variants were found in 12,875 of these regions (MAF >0.05). Genotypes of all rare variants within a gene were collapsed into a single locus, and a global P value per gene region was calculated using Pearson's chi-square test. We repeated the analysis using different MAF thresholds (0.04 , 0.03 , 0.02 , and 0.01). Specifically, we compared the number of cases carrying >1 rare variant minor allele and the number of cases carrying no rare variant minor alleles with the respective numbers for the controls (11,12). In samples with low cell counts, Fisher's exact test was used to assess significance.

Linkage regions were defined as regions with replicating evidence for linkage ($P < 0.05$) in at least 2 independent studies. A total of 11 replicating linkage regions were identified (Table 1), containing 2,056 genes that had >2 rare variants ($\sim 10\%$ of the genome). We constructed a 2×2 contingency table of associated and nonassociated genes (based on different rare variant association P value thresholds) residing within or outside replicating linkage regions. We used Fisher's exact test to test for significant overrepresentation of association signals under linkage peaks. Three definitions of rare variant association signals were used, at significance thresholds of $P = 10^{-8}$, $P = 10^{-5}$, and $P = 10^{-3}$.

RESULTS

A number of rare variants defined by MAF of >0.05 were found to have significant genome-wide associations with RA susceptibility (Figure 1). In addition, we observed an overrepresentation of rare variant signals under RA linkage peaks (Table 2). This relationship was most pronounced when rare variants were defined based on an MAF of >0.05 (overlap $P = 2 \times 10^{-17}$ for signals with association $P = 10^{-5}$), and gradually diminished with reduced MAF thresholds ($P = 0.048$ for MAF >0.04 ; $P > 0.05$ for MAF >0.03 , >0.02 , and >0.01) (data on numbers of linked genes are available from the corresponding author upon request). The reduced level of significance is likely due to lower power with decreasing MAF thresholds and the number of variants considered.

The MHC region is known to be a major risk locus for RA, and multiple alleles of *HLA-DRB1* are known to be strongly associated with disease. We investigated rare variant association signal overlap with linkage peaks excluding the MHC region, as a sensitivity analysis. To remove the effect of *HLA*, we first excluded all genes under the MHC linkage peak from further analysis. As a result, the overlap between rare variant signals (MAF of

0.05) and linkage evidence became nonsignificant ($P > 0.05$) (Table 2). This approach may have resulted in the removal of secondary, non-*HLA-DRB1*, independent signals from analysis. A separate analysis was therefore carried out for genes in the MHC region, by matching case and control samples based on *HLA-DRB1* genotype. This resulted in a much reduced sample size of 594 case-control pairs. We found no significant rare variant associations within this data set.

DISCUSSION

GWAS have identified RA linkage peaks, but followup experiments have typically been unsuccessful in pinpointing the underlying disease risk-enhancing alleles. The majority of fine-mapping studies have focused on common variation, even though linkage scans are more likely to pick up signals generated by more-penetrant rare variants segregating in families. GWAS now offer the possibility to assess (though far from exhaustively) the association between low-frequency or rare variants and disease. This is the first study to evaluate the overlap between genome-wide rare variant signals and replicating linkage peaks for a complex disease.

We observed highly significant overrepresentation of rare variant signals under RA linkage peaks. Although rare variant cluster plots were not scrutinized for the purposes of this study and, therefore, a proportion of association signals may be false-positive, there is no reason to expect differential distribution of these signals across the genome. Our findings suggest the presence of low-frequency RA-associated variants under linkage peaks. This, in turn, indicates that linkage peak fine-mapping experimental designs should explore the full allelic spectrum of sequence variation and should not just be restricted to common polymorphisms.

The significant overlap we detected is largely driven by association signals in the MHC region. This may indicate that additional non-*HLA* RA risk loci remain to be identified at this locus. Failure to detect significant overlap between genome-wide rare variant associations and regions of linkage evidence after accounting for the effect of *HLA-DRB1* was likely due to lack of power in our study, as the sample size diminished substantially after matching. The observed overlap also diminishes with decreasing power at lower MAF thresholds (0.04, 0.03, 0.02, and 0.01).

Current GWAS, including the WTCCC study, have been primarily designed to detect association at common variants and have limited power for the detection of rare variant association signals, as coverage of rare variants on genome-wide SNP arrays is typically poor. Meta-analyses of multiple independent GWAS can boost power substantially by increasing sample size. However, dramatic increases in power will be afforded only by the inclusion of low-frequency or rare variants on commercially available genotyping platforms, or, ideally, through whole-genome sequencing approaches. It remains likely that replicating linkage peaks for complex diseases like RA reflect real disease susceptibility loci of lower frequency than current array resolution allows us to investigate.

Acknowledgments

This study makes use of data generated by the WTCCC. A full list of the investigators who contributed to the generation of the data is available at www.wtccc.org.uk. Funding for the WTCCC was provided by the Wellcome Trust under award 076113.

Drs. Eyre, Barton, Thomson, and Worthington's work was supported by Arthritis Research UK and the Manchester NIHR Biomedical Research Centre. Drs. Lawrence and Zeggini's work was supported by the Wellcome Trust (grants WT088885/Z/09/Z and 079557MA). Dr. Panoutsopoulou's work was supported by grants from Arthritis Research UK (18030) and the Wellcome Trust (WT088885/Z/09/Z).

REFERENCES

1. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. the Wellcome Trust Case Control Consortium (WTCCC). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447:661–78. [PubMed: 17554300]
2. Orozco G, Hinks A, Eyre S, Ke X, Gibbons LJ, Bowes J, et al. the Wellcome Trust Case Control Consortium and YEAR consortium. Rheumatoid arthritis association at 6q23. *Nat Genet*. 2007; 39:1431–3. [PubMed: 17982455]
3. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*. 2007; 316:1336–41. [PubMed: 17463249]
4. 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; 30:1205–13. [PubMed: 2446635]
5. 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; 75:330–7. [PubMed: 15208781]
6. Raychaudhuri S, Thomson BP, Remmers EF, Eyre S, Hinks A, Guiducci C, et al. Genetic variants at CD28, PRDM1 and CD2/ CD58 are associated with rheumatoid arthritis risk. *Nat Genet*. 2009; 41:1313–8. [PubMed: 19898481]
7. Hirschhorn JN. Genomewide association studies: illuminating biological pathways. *N Engl J Med*. 2009; 360:1699–1701. [PubMed: 19369661]
8. 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; 43:30–7. [PubMed: 10643697]
9. Ji W, Foo JN, O’Roak BJ, Zhao H, Larson MG, Simon DB, et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet*. 2008; 40:592–9. [PubMed: 18391953]
10. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type I diabetes. *Science*. 2009; 324:387–9. [PubMed: 19264985]
11. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008; 83:311–21. [PubMed: 18691683]
12. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol*. 2010; 34:188–93. [PubMed: 19810025]
13. Brintnell W, Zeggini E, Barton A, Thomson W, Eyre S, Hinks A, et al. Evidence for a novel rheumatoid arthritis susceptibility locus on chromosome 6p. *Arthritis Rheum*. 2004; 50:3823–30. [PubMed: 15593216]
14. Remmers EF, Plenge RM, Lee AT, Graham RR, Hom G, Behrens TW, et al. STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N Engl J Med*. 2007; 357:977–86. [PubMed: 17804842]
15. 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; 31:315–24. [PubMed: 3358796]

Figure 1. Genome-wide association of rheumatoid arthritis susceptibility with rare variants defined by a minor allele frequency of 0.05. Each point in the Manhattan plot represents a gene region with at least 2 rare variants.

Table 1

Regions with replicating evidence for linkage to rheumatoid arthritis *

Chromosome	Start bp	End bp	No. of genes with 2 rare variants
1	174777835	253296473	423
2	179271727	208640479	118
5	14559233	51169980	108
6	9920074	60851783	485
6	82406588	114880766	112
8	0	19062431	83
12	0	37638862	214
16	7280964	34693695	137
16	42191546	65322493	100
16	60192457	91951399	206
18	36848680	69788988	70

* National Center for Biotechnology Information Build 35 coordinates.

Table 2

Overlap between rare variant (minor allele frequency 0.05) association signals defined by different *P* value thresholds and replicating rheumatoid arthritis linkage peaks

Thresholds for significance *	With MHC [†]		Without MHC	
	No. of genes [‡]	<i>P</i> , with vs. without linkage	No. of genes [‡]	<i>P</i> , with vs. without linkage
<i>P</i> 10 ⁻⁸				
With linkage	8	0.002	7	1.00
Without linkage	8		1	
<i>P</i> 10 ⁻⁵				
With linkage	33	2 × 10 ⁻¹⁷	12	0.28
Without linkage	12		4	
<i>P</i> 10 ⁻³				
With linkage	71	2 × 10 ⁻²⁹	40	0.82
Without linkage	40		5	
Total				
With linkage	2,056		1,883	
Without linkage	10,819		10,735	

* Strength of rare variant association of individual genes.

[†] MHC = major histocompatibility complex.

[‡] Number of "associated" genes with or without linkage evidence compared to the total number of tested genes with or without linkage evidence, as assessed by Fisher's 2-sided exact test.