Review

Genetic epidemiology
Approaches to the genetic analysis of rheumatoid arthritis

Sally John* and Jane Worthington†

*AstraZeneca, Mereside, Alderley Park, Macclesfield, Cheshire, UK.
†ARC Epidemiology Unit, University of Manchester, Stopford Building, Oxford Road, Manchester, UK.

Correspondence: Dr Jane Worthington, ARC Epidemiology Unit, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, UK. Tel: +44 161 275 5651; fax: +44 161 275 5043; e-mail: Jane@fs1.ser.man.ac.uk

Received: 20 November 2000
Revisions requested: 16 January 2001
Revisions received: 27 February 2001
Accepted: 6 March 2001
Published: 27 March 2001

Arthritis Res 2001, **3**:216–220

(Print ISSN 1465-9905; Online ISSN 1465-9913)

Abstract

The basis of susceptibility to rheumatoid arthritis (RA) is complex, comprising genetic and environmental susceptibility factors. We have reviewed the available approaches to the investigation of the genetic basis of complex diseases and how these are being applied to RA. Affected-sibling-pair methods for nonparametric linkage analysis, linkage-disequilibrium-based approaches, transmission disequilibrium testing, and disease-association studies are discussed. The pros, cons, and limitations of the approaches are considered and are illustrated by examples from the literature about rheumatoid arthritis.

Keywords: genetic analysis, rheumatoid arthritis

Introduction

William Heberden in 1806 was probably the first to suggest "some degree of heredity" for rheumatoid arthritis (RA). Subsequent twin- and family-based studies have shown that both genetic and environmental factors influence susceptibility to RA, leading to its description as a complex or multifactorial condition. There have been many attempts to elucidate both the genetic and environmental components, but the aetiology remains largely unclear. In common with many other autoimmune chronic inflammatory conditions, associations with genes of the human leucocyte antigen (HLA) complex have been described. The original observation by Stastny in 1978 [1] of an association with the HLA DRB1 locus in 80 patients has become one of the few examples of a consistently associated gene in RA. Estimates suggest that the HLA locus probably accounts for no more than a third of the total genetic component of susceptibility [2], leaving the majority still to be determined. This review considers some of the approaches currently available for the investigation of the non-HLA genetic basis of susceptibility to RA.

Linkage analysis

Complex diseases do not lend themselves to parametric linkage analysis, as this technique depends on following the inheritance of genetic markers in extended pedigrees to look for cosegregation of marker alleles in affected individuals, under a defined model of inheritance. RA clearly does not follow Mendelian inheritance patterns, and large, extended pedigrees are difficult or impossible to ascertain. Thus, until recently, the search for RA disease genes was targeted to potential candidate genes tested in disease-association studies.

Affected-sibling-pair methods

Nonparametric (model-free) linkage analysis can be carried out on small, nuclear families, such as affected sibling pairs (ASPs), provided sufficient families can be

ASP = affected sibling pair; HLA = human leucocyte antigen; IL = interleukin; RA = rheumatoid arthritis; SNP = single-nucleotide polymorphism; TDT = transmission disequilibrium test.

collected. Technological developments in the early 1990s making possible high-throughput genotyping of informative markers (microsatellites; simple repetitive DNA sequences, highly polymorphic between individuals in terms of the number of repeats) combined with the collection of large numbers of small nuclear families led to groups in Europe [2], Japan [3], the USA [4], and the UK (Worthington J, unpublished data) to embark upon whole-genome screens in search of RA disease genes. This exciting approach of systematically scanning the genome for regions containing disease loci provides the opportunity to identify previously undescribed genes that would never be found by the candidate-gene approach.

The results published to date all represent the first stage of genome screens, in which many potential loci have been identified, but as yet, with the exception of *DRB1*, none has reached the level of statistical significance ($P < 2.2 \times 10^{-5}$) recommended for such approaches [5]. As a result, many of these loci will turn out to be false positives, and only replication studies in independent cohorts will determine the true regions of linkage. Further, the individual studies published so far are limited by a lack of power to exclude regions or to detect loci with modest effects – the likely scenario in RA – and this may require the use of as many as 2000 ASPs [6]. All groups are collecting more families, and plans are already in place to carry out meta-analysis of the data, which may prove to be the most effective way to achieve the necessary power, if the differences between the cohorts are taken into account. We must expect, then, to wait some time before whole-genome screens in RA accurately direct us to disease loci.

On a more encouraging note, even in the initial publications, a number of promising loci have been detected in more than one study. This is perhaps particularly surprising in view of the relative lack of concordance between whole-genome screens reported for some other conditions (e.g. multiple sclerosis, schizophrenia). This lack of agreement has, unfortunately, led to some scepticism about whole-genome screens but actually does not suggest an inherent flaw in the approach: it is more likely a reflection of heterogeneity between the cohorts studied. The possibility of both phenotypic and genetic heterogeneity in complex diseases is a potential difficulty that must be taken into account when attempting to define the genetic basis of a complex disease, whether using linkage- or nonlinkage-based approaches.

Establishing evidence of linkage using ASP methods is just the first step, and one of the greatest challenges to molecular genetics is to identify a disease gene from a region of linkage. The size of an initial linkage region may be many tens of centimorgans, and recent publications have shown that chance variation in the location estimate is substantial [7,8], suggesting that it may be necessary to follow up a large area on either side of the linkage peak. Experience so far suggests that even with fine mapping using a dense marker map, it has not been possible to narrow down regions to less than 10 cM in any complex disease. As regions of this size may contain hundreds of genes, it is vital to define a smaller region by linkage-disequilibrium mapping before moving on to target genes.

Linkage-disequilibrium analysis and association studies

Family-based association tests

A large area of linkage may be narrowed down by using methods that detect association in the presence of linkage. As association between a marker and a disease locus depends upon the presence of linkage disequilibrium, in an outbred population, association can be detected only over a small distance, typically less than 1 cM. A number of family-based association methods have been proposed, the most commonly used being the transmission disequilibrium test (TDT) [9].

The test examines the transmission of potential disease alleles from a parent who is heterozygous for the marker to an affected offspring. It is a test of association only in the presence of linkage, and because family members act as controls, spurious associations due to population differences do not arise. The original test uses a single affected offspring and both parents. A number of extensions to the original TDT have been proposed that allow both parents and unaffected sibling to be analysed, making maximum use of incomplete nuclear family data. It is now also possible to analyse dichotomous and quantitative variables (e.g. age at onset) and to include covariates (e.g. exposure to smoking) in the analysis. An extensive list of available methods and software can be found at the Genetic analysis web page at Rockefeller University (http://linkage.rockefeller.edu/).

Tests are available for both single-nucleotide polymorphisms (SNPs) and multiallelic markers such as microsatellites. As the TDT is dependent upon the number of informative transmissions, microsatellite markers are often more useful. If 100 parents are genotyped, 80 informative transmissions will be expected for a microsatellite marker with 80% heterozygosity, whereas the maximum heterozygosity measure for a SNP is only 50%. However, at least two multilocus haplotype methods have been developed for the TDT, which should overcome the low information content of single SNPs by combining up to four SNPs in a single haplotype.

At present, there are few published studies in RA using TDT methods. Recent papers have looked at regions of linkage and used the TDT to test for association, leading to more significant results than were observed using linkage [10,11]. The TDT has become a very versatile methodology, allowing a range of family structures, marker types, and phenotype data to be analysed simultaneously. This versatility, coupled with the increased power of an association-based method, will inevitably lead to increased use of the TDT in the search for RA susceptibility genes.

Case–control studies

Linkage studies are resource intensive and dependent on the availability of large family collections. It is therefore not surprising that many investigators have chosen to target candidate genes directly. These studies are usually association based, using case–control cohorts. A number of polymorphisms in genes thought to be involved in RA pathology have been investigated, but results have often been conflicting. An example is the interleukin(IL)-1 gene cluster, containing IL-1B, IL-1A, and the IL-1-receptor antagonist. A number of studies [e.g. 12] have suggested that polymorphisms in this gene cluster are associated with RA, whereas others have shown no association [e.g. 13]. This apparent inconsistency may be explained by a number of factors, including clinical heterogeneity (associations are often only shown with certain subsets of disease), genetic heterogeneity (it should not be unexpected to find ethnic differences in associations), and study design (small, underpowered sample sizes, poor quality control of genotyping data, and inappropriate selection of controls will all contribute to inconsistent findings).

Despite the potential difficulties, association studies have the significant advantage over linkage studies of having greater power to detect small effects. For example, only 123 affected individuals in a case–control study would be required to detect a genetic relative risk of 2 for a disease allele with a frequency of 10% with 80% power, at $P=0.05$ [6]. Although the genetic relative risk associated with an unknown disease gene cannot be established accurately, there is much evidence that no disease gene in RA will have an effect greater than HLA and that some RA genes may well have a genetic relative risk no greater than twofold. In addition to offering increased power, the move towards genotyping SNPs rather than microsatellites means that case–control studies are more efficient.

Even when reasonable sample sizes are used, applying a significance level of $P=0.05$ will still lead to 1 in 20 results being false positives. In order to minimize type 1 error, it would be desirable to design studies with sufficient power to detect an effect at a level of significance corrected for the number of markers or genes to be tested. The obvious drawback of this rigorous approach is that this sort of correction for multiple tests will result in unrealistic sample sizes. For this reason, it is perhaps more appropriate to accept a P value of <0.05 in an initial study and to replicate the result in an independent data set [14].

The selection of appropriately matched controls has also been the subject of discussion within the community of geneticists. Ethnically unmatched controls may lead to positive results due to population stratification. If two populations have subtle genetic differences and the cases come predominantly from one population, positive associations will be observed but the true association will be with the population rather than the disease. In reality, the extent of this problem is unclear, because false-positive results occur for many reasons. The problem has recently been addressed by Pritchard and Rosenberg [15], who propose using a panel of unassociated markers to test for population stratification within the cohort under investigation.

Future considerations

The rapid pace of developments in molecular genetics and molecular medicine make it almost impossible to accurately predict more than a couple of years into the future. It seems likely that our investigation of RA genetics will continue, in the immediate future, to be based on a combination of linkage and association studies, with refinements to improve the power and sensitivity. Linkage mapping of ASP collections will probably use a higher density of markers, and information derived from other sources such as animal models may be used to target the linkage studies. The investigation of loci homologous to regions mapped in rodent models of disease has proved fruitful in a number of diseases, including arthritis [16]. With the human genome now sequenced and the mouse sequence expected within the year, the accurate targeting of homologous regions for linkage analysis will be greatly facilitated.

The ASP collections may also become the samples of choice for association-based studies. Case–control and TDT methods have routinely used sporadic cases, but in a complex disease such as RA, any study design based on sporadic cases may be selecting more for environmental than for genetic factors. Risch [17] has advocated the use of ASPs in an association study design with unrelated controls as the most powerful approach to detect disease genes. For example, using 102 ASPs (408 individuals in total) has an 80% power to detect a genetic relative risk of 2 (for the heterozygote) for a disease allele of 20% frequency with a significance level of $P=5\times10^{-8}$. Intuitively, taking cases with a family history should increase the chances of detecting a genetic effect. A recent publication demonstrated an association to the tumour necrosis factor receptor II in two independent data sets in which the case had a family history of disease (defined as at least one affected first-degree relative) [18]. This association was not observed in a cohort of sporadic cases.

So far, association studies in RA have concentrated on a few, well established candidate genes. In theory, it is possible to search the whole genome by association methods. The likelihood of success using this approach depends upon the number of markers typed and the extent of linkage disequilibrium that exists between markers. It is beyond the scope of this review to discuss whole-genome linkage-disequilibrium mapping and whether it will be applied to RA. More thorough discussion of the feasibility of this approach has recently been published [17,19].

With the completion of the human genome sequence, all transcribed genes should soon be identified. Additional information about tissue expression and functional domains will allow us to make much more educated decisions about which genes to target. With initiatives such as the SNP consortium releasing >300,000 SNPs into the public domain, it will soon be possible to select SNPs in candidate genes from a list of all transcribed genes in the genome. This more focused approach may lead to greater success in detecting disease genes, because testing potentially functional SNPs within genes for association decreases the dependence upon linkage disequilibrium. However, even for a single gene, there is no clear consensus about how many SNPs one might need to analyse. A recent publication examining SNPs around the *APOE* locus failed to find an association with the majority of the common SNPs within a 1.5-Mb region of the gene [20]. Case–control studies have most commonly been used to look at single markers; analysing several markers within a gene or small region has been more problematic, because it is difficult to determine haplotypes in the absence of family information. There is now a concentrated effort to evaluate methods of haplotyping unrelated individuals, and a recent publication successfully identified the *APOE* locus using haplotyping methods [21]. Drysdale *et al* used haplotypes in the β_2-adrenergic receptor to detect association with drug responsiveness; they suggested that haplotypes were more successful in detecting associations than in analysing individual SNPs [22].

Conclusion

Having reached the landmark event of sequencing the human genome, perhaps we are now in a position to really begin dissecting the aetiology of RA. Ultimately, this will be achieved only by using a combination of the techniques described in this review, together with high-quality phenotypic and epidemiological data. This will also require the development of methods of analysis based on more sophisticated models of complex disease which allow for gene–gene and gene–environment interactions.

References

1. Stastny P: **Association of the B-cell alloantigen DRw4 with rheumatoid arthritis.** *N Engl J Med* 1978, **298**:869–871.
2. Cornelis F, Faure S, Martinez M, Prud'homme JF, Fritz P, Dib C, Alves H, Barrera P, de Vries N, Balsa A, Pascual-Salcedo D, Maenaut K, Westhovens R, Migliorini P, Tran TH, Delaye A, Prince N, Lefevre C, Thomas G, Poirier M, Soubigou S, Alibert O, Lasbleiz S, Fouix S, Weissenbach J: **New susceptibility locus for rheumatoid arthritis suggested by a genome wide linkage study.** *Proc Natl Acad Sci USA* 1998, **95**:10746–10750.
3. Shiozawa S, Hayashi S, Tsukamoto Y, Goko H, Kawasaki H, Wada T, Shimizu K, Yasuda N, Kamatani N, Takasugi K, Tanaka Y, Shiowzawa K, Imura S: **Identification of the gene loci that predispose to rheumatoid arthritis.** *Internat Immunol* 1998, **10**: 1891–1895.
4. Jawaheer D, Seldin MF, Amos CI, Chen W, Montiero J, Criswell L, Albani S, Nelson L, Clegg DO, Pope R, Shroeder HW, Bridges SL, Pisetsky DS, Kastner D, Wilder R, Pincus T, Callahan L, Gregersen PK: **A genome wide screen for allele sharing in the first 300 sibling pairs of the NARAC collection.** *Arthritis Rheum* 2000, **43(suppl)**:S390.
5. Lander E, Kruglyak L: **Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results.** *Nature Genet* 1995, **11**:241–247.
6. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273**:1516–1517.
7. Roberts SB, MacLean CJ, Neale MC, Eaves LJ, Kendler KS: **Replication of linkage studies of complex traits: an examination of variation in location estimates.** *Am J Hum Genet* 1999, **65**:876–884.
8. Clerget-Darpoux F, Selinger-Leneman H, Babron MC: **Why are genome scans of multifactorial diseases so discordant.** *Genet Epidemiol* **3**:241.
9. Spielman RS, McGinnis RE, Ewens WJ: **Transmission test for linkage disequilibrium: the insulin gene region and insulin dependent diabetes mellitus (IDDM).** *Am J Hum Genet* 1993, **52**:506–516.
10. Myerscough A, John S, Barret JH, Ollier WER, Worthington J: **Linkage of rheumatoid arthritis to insulin-dependent diabetes mellitus loci: evidence supporting a hypothesis for the existence of common autoimmune susceptibility loci.** *Arthritis Rheum* 2000, **43**:2771–2775.
11. Fife S, Fisher SA, John S, Worthingon J, Shah CJ, Ollier WER, Panayi GS, Lewis CM, Lanchbury JS: **Multipoint linkage analysis of a candidate gene locus in rheumatoid arthritis demonstrates significant evidence of linkage and association with the corticotropin-releasing hormone genomic region.** *Arthritis Rheum* 2000, **43**:1673–1678.
12. Cantagrel A, Navaux F, Loubet-Lescoulie P, Nourahashemi F, Enault G, Abbal M, Constantin A, Laroche M, Mazieres B: **Interleukin-1 beta, interleukin-1 receptor antagonist, interleukin-4, gene polymorphisms: relationship to occurrence and severity of reheumatoid arthritis.** *Arthritis Rheum* 1999, **42**:1093–1100.
13. Bailly S, Hayem G, Fay M, Kahn MF, Gougerot-Pocidalo MA: **Absence of correlation between IL-1 alpha intron 6 polymorphism and arthritis.** *Br J Rheumatol* 1995, **34**:1123–1126.
14. Nistico L, Buzzetti R, Pritchard L, Van der Auwera B, Giovannini C, Bosi E, Larrad MT, Rios MS, Chow CC, Cockram CS, Jacobs K, Mijovic C, Bain SC, Barnett AH Vandewalle CL, Schuit F, Gorus FK, Tosi R, Pozilli P, Todd JA: **The CTLA4 gene region of chromosome 2q33 is linked to and associated with type I diabetes.** *Hum Mol Genet* 1996, **5**:1075–1086.
15. Pritchard JK, Rosenberg NA: **Use of unlinked genetic markers to detect population stratification in association studies.** *Am J Hum Genet* 1999, **65**:220–228.
16. Barton A, Eyre S, Myerscough A, Silman A, Ollier W, Worthington J: **A novel RA susceptibility locus on chromosome 17 identified using syntenic mapping approaches.** *Arthritis Rheum* 2000, **43(suppl)**:S271.
17. Risch N: **Determining genetic variants in the new millenium.** *Nature* 2000, **405**:847–856.
18. Barton A, John S, Ollier WER, Silman A, Worthington J: **Association between rheumatoid arthritis and polymorphism of the tumour necrosis factor 2 (TNFR2) but not TNFR1 in caucasions.** *Arthritis Rheum* 2001, **44**:61–65.
19. Weiss KM, Terwilliger JD: **How many diseases does it take to map a gene with SNPs?** *Nature Genet* 2000, **26**:151–157.
20. Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J, Finch KL, Stevens JF, Livak KJ, Slotterbeck BD, Slifer SH, Warren LL, Conneally PM, Schmechel DE, Purvis I, Pericak-Vance MA, Roses AD, Vance JM: **SNPing away at complex disease: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease.** *Am J Hum Genet* 2000, **67**:383–395.
21. Fallin D, Cohen A, Essioux L, Chumakov, I, Blumenfeld M, Cohen D, Schork N: **Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease.** *Genome Res* 2001, **11**:143–151.

22. Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, Liggett SB: **Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness.** *Proc Natl Acad Sci USA* 2000, **97**:10483–10488.