Genetic Studies of Rheumatoid Arthritis: Progress and Challenges

Changsoo Kang
Department of Biology, Research Institute of Basic Sciences, College of Natural Sciences, Sungshin Women’s University, Seoul, Korea

Rheumatoid arthritis (RA) is a systemic inflammatory disease associated with both genetic and environmental factors. The \textit{DRB1} gene at the human leukocyte antigen (HLA) locus of chromosome 6p21.3 was the first genetic factor associated with RA to be identified in the 1980s; however, identification of causative genes other than those at the HLA locus has been challenging for geneticists because of the strong linkage disequilibrium in this locus and the non-Mendelian inheritance pattern of RA. Recent advances in high-throughput single nucleotide polymorphism genotyping technologies and bioinformatic analysis tools have facilitated the identification of positive associations of hundreds of genes with RA using family-based linkage analyses and genome wide association studies. Some of the RA associated genes at non-HLA loci are as follows: \textit{PADI4}, \textit{PTPN22}, \textit{STAT4}, and \textit{TNFAIP3}. In this paper, we describe the pathological mechanisms mediated by these genes. In addition, we review results of previous genetic studies of RA and future challenges in connecting the dots of missing heritability in the post-genome-wide association study era. (J Rheum Dis 2015;22:274-281)

Key Words. Rheumatoid arthritis, Genetics, Mutation, Linkage, Association

INTRODUCTION

Rheumatoid arthritis (RA, OMIM#180300) is a systemic inflammatory disease that causes progressive joint destruction, particularly of the feet and hands. The prevalence rate of RA worldwide is about 0.5% to 1% in adults and the rate of occurrence is higher in females than in males. Both environmental and genetic factors are known to be responsible for the susceptibility to RA and also contribute to the phenotype of this disorder [1]. Environmental factors include geographical area, gender, infectious agents, and lifestyle factors such as smoking and diet [2,3]. Many studies have been undertaken to investigate the biological mechanisms underlying RA; however, its etiology is still unclear. Previously, RA was classified as an autoimmune disease because the rheumatoid factor (RF) in the serum of RA patients binds to the Fc portion of immunoglobulins. The estimated percentage of individuals seropositive for the self-reactive RF was about 80% of the affected individuals. The RF serves as an initiator in the pathogenesis of immune complex-mediated disease by predisposing patients to a more aggressive and destructive form of RA [4]. However, RF presents in the serum of other autoimmune diseases as well; therefore, it may not play a specific role in RA as previously postulated. Notably, another autoantibody known as the anti-citrullinated protein antibody (ACPA) has a higher specificity in RA patients suggesting that citrulline may be the key antigenic determinant in the pathogenesis of RA [5]. Although the functional roles of these autoantibodies in the pathogenesis of RA are poorly understood, it is proposed that RF and ACPA form immune complexes in patients with RA. These immune complexes then influence the release of chemotactic factors and recruitment of inflammatory cells to the joints along a chemotactic gradient, where they are activated and con-
Genetic approaches can be useful tools for investigating the etiology of RA. In cases where genetic factors are co-segregated with RA, the use of genetic tools may facilitate identification of the causative genes for RA. Considerable evidence exists to suggest that genetic factors may play a role in the pathogeneses of RA. Heritability of RA is about 53% to 60%, and the relative risk of developing RA is 2% to 17% in siblings (\( \lambda_s \)) of RA patients, 5% to 10% in same sex dizygotic twins, and 12% to 30% in monozigotic twins [6]. In addition, family clustering of RA has been reported. Many researchers have used these data on the role of genetic factors in the pathogenesis of RA as the basis for genome-wide linkage and association studies to identify the causative factors of this disorder. Using this approach, multiple studies have successfully identified genetic causes of RA. In this study, we describe the recent genetic studies on RA and the challenges that remain to be overcome.

**GENETIC STUDIES OF RHEUMATOID ARTHRITIS**

**Human leukocyte antigen locus**

The human leukocyte antigen (HLA) locus (human major histocompatibility complex, MHC) on the human chromosome 6p21 is the most heterogeneous and gene-rich region of the human genome. There are about 224 genes in this 3.6-Mb region, and 40% of them are known to be associated with immune system functions. In 1969, it was first reported that genetic factors in the HLA region are associated with RA [7] and Stastny [8] in 1978 observed that the frequency of the B-cell alloantigen HLA-DRw4 was 70% in Caucasian patients with erosive seropositive RA compared to 28% in controls (p-value < 10\(^{-5}\)). They hypothesized that RA is associated with genetic factors in the HLA-D region and that immunogenic factors linked to HLA are involved in RA pathogenesis. Other studies on distribution of DR antigens confirmed this association of DR4 with RA and reported that the overall relative risk of developing RA in DR4 homozygotes was 124.46 [9]. Other DR antigens such as DR1, DR2, and DR3 are also known to increase the risk for developing RA; however, their relative risks are moderate compared to that of the DR4 allele. However, DR5, DRw6, and DR7 may protect against RA.

**Shared epitope hypothesis**

Although the early studies on RA described above reported for the first time that the genomic region of HLA-DR is associated with RA however, the specific gene(s) causing RA was still unknown. This region harbors two functional \( \beta \) chain genes, named DR\( \beta \)1 and DR\( \beta \)III that are tightly linked with the DQ locus thereby constituting haplotypes that co-segregate as a block. Therefore, identifying a single gene associated with RA has been a significant challenge for researchers. Further, intensive sequencing effort in the expanded DQ and DR regions revealed that most of the variability in haplotypes emerging from the HLA-DQ to HLA-DR regions was due to polymorphisms in the third hypervariable region of the DRB1 gene encoding for the \( \beta \) chain of the DR molecule. Further, it was shown that several HLA-DRB1 alleles associated with RA share similarities in amino acid sequences (\( ^7^0^Q^R^R^A^A^\), \( ^7^0^R^R^A^A^\), or \( ^7^0^K^R^R^A^A^\)) in the epitope-recognition region of this protein. This is the basis of the shared epitope (SE) hypothesis proposed by Gregersen et al. [10] in 1987, which suggests that this locus accounts for one third of the genetic susceptibility to RA. Further conditional and haplotype analysis revealed that amino acid positions of 11, 71 and 74 in the HLA-DRB1 are the major contributors to the seropositive RA susceptibility and associated with RA severity, mortality and tumor necrosis factor (TNF) inhibitor treatment response [11,12].

The association of DRB1 SE alleles with RA susceptibility varies among different ethnic groups. Genetic studies of African Americans and Hispanic populations show that SE alleles are not associated with susceptibility to RA [13,14]. In addition, DRB1*0401 and *0404 alleles show significant association with RA in Caucasian patients while these alleles are not associated with RA in Asian patients [15,16]. In contrast, DRB1*0405 and the non-SE allele *0901 are strongly associated with an increased risk of RA in East Asian populations [15,16]. In particular, the genotype DRB1*0405/*0901 shows the most significant association with RA [15].

**Genetic factors other than DRB1 in the HLA locus**

Several studies have demonstrated a positive association between DRB1 SE alleles and RA. The SE hypothesis, however, cannot explain all the genetic factors associated with RA susceptibility. This implies that there might be other causative genes besides the DRB1 gene either in the HLA locus or in other chromosomal loci. Several association studies to identify genetic markers associated with
RA in unrelated RA cases and controls have reported a positive association of two intronic single nucleotide polymorphism (SNP) markers in the major histocompatibility complex class I chain-related gene A (MICA, p-value=0.068) and major histocompatibility complex class II, DQ beta 2 (HLA-DQB2, p-value=0.012) genes with RA susceptibility independent of the HLA-DRB1 SE alleles [17]. Okamoto et al. [18] in 2003 reported that the T allele of a SNP (rs96452, p-value=0.0062) in the promoter region of the IkBL gene encoding the inhibitor of kappa light chain gene enhancer in B cells-like may increase RA susceptibility by disrupting the normal binding motif for the transcriptional repressor δ EF1. However, the associations of the variants in the MICA, HLA-DQB2, and IkBL genes were only statistically normal and it is therefore possible that the positive associations are due to linkage disequilibrium (LD) with the DRB1 gene. Therefore, further investigations are needed to reveal the functional role of these variants along with studies to identify additional genetic factors other than the DRB1 gene in HLA locus. These investigations and studies could prove to be tremendously challenging for geneticists.

**Family-based linkage analysis**

The risks for developing several disorders are mostly both genetic and environmental factors. As the effect of genetic factor(s) especially single genes is stronger than environmental factors there is a higher chance of disorders clustering in families. Genome-wide linkage analysis and subsequent sequencing of candidate genes at linkage intervals have generally been regarded as useful approaches to identify genetic causes of RA in families with multiple affected members. Since the 1980s, several genome-wide linkage studies have been performed in multicase RA families and significant linkage of RA was consistently observed with the markers in chromosome 6p21 harboring the DRB1 gene [19-26]. In 1998, Cornélis et al. [20] recruited 97 European Caucasian nuclear families and found significant linkage of the markers on HLA region in chromosome 6p21 (p-value<2.5×10^{-5}), and nominal linkage of 19 markers in 14 other regions (p-value<0.05). In addition, this study also proposed that CD80 and CD86 were candidate genes for RA (Table 1). These genes are known to be involved in antigen-specific T cell recognition however, intensive sequencing of other candidate genes in the linkage loci to identify causative mutations remains to be performed. Another linkage study found nominal linkage of the markers on chromosome 2q35 and suggested that NRAMP1 may be a candidate gene for RA. However, the maximum logarithm of odds (LOD) score for NRAMP1 was not high enough to support significant linkage (LOD score=1.01) [23].

The results of the individual linkage studies described above (Table 1) could not be replicated by other studies because of insufficient power to detect linkage except that of the well-established HLA susceptibility locus. This failure to detect significant linkage for most genes other than the HLA locus may be because of the followings: all linkage studies were done in small or moderately sized nuclear families in which only the causative gene shared most families such as DRB1 gene is detectable. Since each of these families might not be affected by a deficit in the same gene, using a single large family with at least 3 generations might potentially result in successful linkage analysis because it can be assumed that affected individuals in a family would share the same genetic cause for RA. In addition, RA is a complex trait that does not typically follow Mendelian inheritance pattern because of the low penetrance of this disorder. The observation that 15% of individuals carrying the DRB1 *0405 or *0901 alleles in either homozygous or heterozygous form are phenotypically normal [15] supports the idea of low penetrance of the risk alleles of RA. Thus, this ambiguous inheritance pattern of RA prevents identification of genetic causes by typical linkage approach. Genetic linkage or next-generation sequencing studies in consanguineous families showing recessive inheritance of RA and high disease penetrance might be promising strategies to circumvent this difficulty.

**Linkage disequilibrium mapping and PADI4**

The development of high-throughput SNP genotyping together with bioinformatics analysis platform enabled researchers to perform genome-wide association study (GWAS) as an alternative to linkage study. GWAS compares the SNP frequencies in unrelated cases and normal control subjects at the genome-wide level by genotyping for several thousand SNPs and then identifies alleles shared predominantly by cases. If one type of SNP allele is statistically significant in the unrelated cases, the SNP is regarded to be associated with the phenotype. Usually, the associated SNP is not disease causing, in contrast, it is the marker that co-segregate to the progeny with the true mutation as LD block in the unrelated population.

GWAS was believed to be effective at detecting common variants with weak genetic effects: this is common in...
Table 1. Notable genetic studies of rheumatoid arthritis

<table>
<thead>
<tr>
<th>Reference</th>
<th>Method</th>
<th>Summary of finding</th>
<th>Gene or chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>[8]</td>
<td>Serological testing in 80 RA patients and controls</td>
<td>B-cell alloantigen HLA-DRw4 were statistically more frequent in patients (70%) than in normal controls (28%)</td>
<td>6p21.3</td>
</tr>
<tr>
<td>[10]</td>
<td>DNA sequencing</td>
<td>Found association of DRB1 SE alleles with RA</td>
<td>HLA-DRB1</td>
</tr>
<tr>
<td>[17,18]</td>
<td>Case and control study by genotyping SNPs in MHC locus</td>
<td>Suggested nominal association of the intronic and promoter SNPs with RA independently of DRB1 SE</td>
<td>MICA, HLA-DQB2, IKB</td>
</tr>
<tr>
<td>[11,12]</td>
<td>GWAS</td>
<td>Identified three major causative variants in DRB1 with RA</td>
<td>HLA-DRB1</td>
</tr>
<tr>
<td>[19-26]</td>
<td>Family-based linkage analysis</td>
<td>Identified significant linkage of RA to the markers in several chromosomes</td>
<td>1p36, 2q33, 2q35, 3q13, 6p21.3, 14, 16p13-q12.2, 18q22-23</td>
</tr>
<tr>
<td>[27]</td>
<td>Association study of candidate region (1p36)</td>
<td>Revealed functional PADI4 haplotypes are associated with RA by increasing autoantibodies against citrullinated peptides</td>
<td>PADI4</td>
</tr>
<tr>
<td>[29]</td>
<td>Association study of candidate region (2q33)</td>
<td>Found association of haplotype in the intronic region of STAT4 with both RA and SLE</td>
<td>STAT4</td>
</tr>
<tr>
<td>[34]</td>
<td>Association study of candidate genes</td>
<td>Identified a missense mutation in PTPN22 and found risk allele increase systemic autoimmunity</td>
<td>PTPN22</td>
</tr>
<tr>
<td>[37]</td>
<td>GWAS</td>
<td>Found an intergenic SNP associated with RA</td>
<td>TNFAIP3</td>
</tr>
<tr>
<td>[40]</td>
<td>GWAS</td>
<td>Identified 7 risk loci for RA in the Caucasian population</td>
<td>IL6ST, SPRED2, RBPI, CCR6, IRF5, PXK, IL2RA, CCL21, AFF3</td>
</tr>
<tr>
<td>[39]</td>
<td>GWAS</td>
<td>Identified 9 loci associated with RA in the Japanese population</td>
<td>B3GNT2, ANXA3, CSF2, CDB3, NFKBIE, ARID5B, PDE2A-ARAP1, PLD4, PTPN2</td>
</tr>
</tbody>
</table>


complex traits including RA. Suzuki et al. [27] in 2003 used this as the basis of LD mapping (although it was not a full GWAS) by genotyping 118 SNPs distributed over the previously known linkage interval of chromosome 1p36 in 1,566 Japanese unrelated cases of RA and controls. This study showed that four exonic SNPs (p.G55S, p.V82A, p.G112A, and p.L117L) constituting the functional haplotypes of PADI4 encoding peptidylarginine deiminase, are associated with RA (p-value=0.000008). Further, functional analysis revealed the role of the PADI4 gene in RA pathogenesis by showing that the susceptible haplotype increases the stability of the PADI4 transcript leading to production of higher levels of citrullinated peptides serving as autoantigens (ACPA) in sera from RA patients (Figure 1) [27]. The mechanism of increase in autoantibodies against citrullinated proteins (ACPA) upon the replacement of arginine with citrulline on the proteins by PADI4 is not fully understood. However, these results are highly convincing considering that citrullinated proteins are known to be epitopes recognized by RA specific autoantibodies. In addition, the genetic contribution of PADI4 haplotypes to the RA susceptibility shows ethnic differences with large effects in the East Asian populations such as Japanese and Koreans; however, no positive association in Caucasians from United Kingdom, France, and Spain have been reported. The cause of this ethnic difference in association was highlighted by the finding that PADI4 risk haplotype predis-

![Figure 1. Conversion of arginine residue to citrulline by peptidyl arginine deiminase, type IV (PADI4).](image-url)
poses smokers to RA; thus, different smoking prevalences in the populations might be the cause of the ethnically different association of \textit{PADI4} with RA [28]. Thus, genetic association studies in unrelated cases and controls to identify causes of RA established \textit{PADI4} as one of the first susceptibility genes for RA, apart from the HLA region.

**Linkage disequilibrium mapping and \textit{STAT4}**

Significant linkages of the SNP markers in the chromosome 2q33 region to RA have been reported previously in the study with 642 Caucasian families [26]. Fine LD mapping of this linkage region encompassing 50-Mb, revealed that a SNP haplotype in the third intronic region of signal transducer and activator of transcription 4 (\textit{STAT4}) was associated with increased susceptibility to both RA and systemic lupus erythematosus (SLE) [29]. The risk for RA in individuals carrying two copies of the risk haplotypes was 60% higher than in those carrying no copies of the risk haplotypes. This association was reported in multiple populations including Swedish, Spanish, Dutch, Korean, and Japanese populations [30-32]. \textit{STAT4} is a cytosolic transcription factor that is transported into the nucleus after activation by JAK-mediated phosphorylation. This activated \textit{STAT4} together with interferon-\(\gamma\), interleukin (IL)-12, and IL-23 induces naïve CD4\(^+\) T cells to differentiate to Th1 and Th17 cells. These Th1 and Th17 helper T cells are known to be associated with chronic inflammatory disorders [33]. Further, disruption in the normal functions of \textit{STAT4} caused resistance against autoimmune disease in a mouse model and in ameliorate phenotypes similar to RA. Therefore, \textit{STAT4} may be a useful therapeutic target for treatment of autoimmune disease including RA.

**Candidate gene approach and \textit{PTPN22}**

Begovich et al. [34] in 2004 found a significant association of missense variants with RA by genotyping and comparing the frequencies of 87 putative functional SNPs localized at RA candidate genes and linkage regions. This association study revealed that substitution of arginine with tryptophan at position 620 in the hematopoietic specific protein tyrosine phosphatase encoded by \textit{PTPN22} increased RA susceptibility. The frequency of the risk allele encoding tryptophan was much higher in cases (28%) than in controls (17%) in Caucasians. Further, this allele was suggested to disrupt the P1 proline-rich motif that is involved in interaction with c-Src kinase thereby leading to alterations in its normal functions such as the negative regulation of T-cell activation. Generation and characterization of knock-in mice carrying the analogous mutation PEST domain phosphatase p.R619W, showed expansion of effector T and B cell populations and increased autoantibody production resulting in a loss in self-tolerance and autoimmunity [35]. This functional SNP was not only associated with RA; positive associations with other immune diseases including SLE, type 1 diabetes, and Graves’ disease were also reported [36]. Notably, this variant is only found in Caucasian and Hispanic populations; however, the risk allele frequencies in East Asian and African populations were zero. Thus, this is additional example of ethnically different association of causative genes with RA susceptibility [34].

**GWAS and \textit{TNFAIP3}**

The genetic association of \textit{PADI4} and \textit{PTPN22} with RA was discovered by LD mapping in the linkage region and candidate gene approach in unrelated RA cases and controls, respectively. The first association study at a genome-wide level was performed by the Wellcome Trust Case Control Consortium in 2007 [37]. This GWA study confirmed the association of SNPs within the HLA locus and the \textit{PTPN22} with RA and also screened 9 SNPs associated with RA by genotyping of 1,860 cases and 2,938 controls for the SNPs at a genome-wide level. The association of one intergenic SNP located near tumor necrosis factor-\(\alpha\)-induced protein 3 (\textit{TNFAIP3}) was replicated in the independent sample group consisting of 5,063 cases and 3,849 controls. \textit{TNFAIP3} encodes an A20 protein, which is the negative feedback regulator of nuclear factor-\(\kappa\)B signaling in response to proinflammatory stimuli. Knockout mice for \textit{Tnfaip3} showed spontaneous development of a destructive polyarthritis and many other phenotypes similar to those of RA [38]. This mouse data supports the hypothesis that \textit{TNFAIP3} is associated with RA.

**Future challenges for the genetic studies of RA**

Genetic studies of RA using linkage analysis, candidate gene approach, and GWAS were regarded as powerful tools in identifying causative genes or DNA markers associated with RA susceptibility. In particular, GWA studies based on high-throughput SNP genotyping technologies facilitated the identification of several unknown genetic causes of RA [39-41]. Currently 98 biological candidate genes at 101 chromosomal loci are known to be associated with RA [41]. However, there are several chal-
lenges that future studies have to address. First, positive associations of these candidate genes with RA susceptibility were based on genotyping pre-selected SNPs with relatively high frequency in commercially available SNP chips, and it is unclear whether these associated SNPs have direct functional effects on RA susceptibility. Thus, it is possible that other neighboring SNPs in the same or other genes that are in LD with the associated SNPs are true disease causing genetic factors. Therefore, intensive sequencing efforts to reveal functional variants in candidate genes in a large number of RA cases are needed. Secondly, the most commonly associated SNPs with RA revealed by GWAS are common variants that could only explain ∼36% of the overall disease liability or 65% of the total heritability for developing RA [41,42]. Therefore, half of the total heritability remains unexplained. The effects of rare causative variants with frequencies less than 1% might explain this missing heritability because conventional GWAS cannot detect the effect of these rare variants. However, recently emerging whole-exome or genome sequencing technologies might be promising approaches to detect rare causative variants for RA to explain some portions of the missing heritability.

Clinical application of genetic data in RA diagnosis and treatment

The final goals in the genetic studies RA are to predict susceptibility, progressivity, and severity of this disorder and to prescribe suitable medicine to the RA patients. Although, genetic heterogeneities of RA might be critical obstacles in the application of genetic data for clinical usage, there are few genetic data that can predict disease severity and responses to treatments. One example is the recent finding that amino acids at positions 11, 71, and 74 of HLA-DRB1 are associated with RA severity and mortality [12]. Therefore, the RA patients carrying genotypes that contributing to increase RA severity may be expected to experience rapid disease progression. Thus, more extensive and rapid use of therapeutic agents might be helpful to these individuals. Another clinical use of genetic data in RA is to predict extent of drug efficacy in the individuals when they are treated with biologics such as TNF blockers, because 20% to 30% of the RA patients do not respond to anti-TNF antibodies. If individual’s responses to these therapies can be predicted, suffering from side effects and unnecessary cost expenditure can be avoided. However, these clinical applications to predict RA severity and drug responses need to be designed very carefully due to that single genetic factor cannot predict them sufficiently. Thus, additional genetic data for RA severity and treatment together with environmental factors are pre-requisites for proper clinical application of RA genetic data.

CONCLUSION

Family-based linkage analysis and GWASs have revealed hundreds of genes associated with RA susceptibility. Epitope recognition region of the HLA-DRB1 gene was initially known to be strongest genetic causative agent of RA; however, many other associated genes have now been identified with the help of new genetic technologies. Although, there are still some heritability factors that cannot be explained by the effects of only the reported genes, sequencing in large number of RA patients using exome or genome sequencing technologies might shed light on novel genetic associations of RA by identifying variants with low allele frequency. In addition, epigenetic analyses to reveal genome-wide methylation or histone modification patterns may be a useful approach to reveal interaction of genes and environment. However, the current information on the genetic architecture of RA remains to be further characterized to be applicable in disease diagnosis and treatment.

ACKNOWLEDGMENTS

This study was supported by the Sungshin University Research Grant of 2015.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

5. Schellekens GA, de Jong BA, van den Hoogen FH, van de Putte LB, van Venrooij WJ. Citrulline is an essential constituent of antigenic determinants recognized by rheuma-
36. Zheng J, Ibrahim S, Petersen F, Yu X. Meta-analysis reveals an association of PTPN22 C1858T with autoimmune dis-
Genetic Studies of RA


