Data S5

**Rheumatoid arthritis**

Rheumatoid arthritis (RA, MIM 180300) is a systemic, chronic inflammatory disorder whose root cause is unclear. The clinical hallmark of RA is an inflammatory arthritis with a predilection for specific diarthrodial (freely movable) joints. The clinical course of RA is extremely variable: some patients suffer mild, self-limiting arthritis while others develop progressive multi-system inflammation with profound morbidity and mortality. Auto-antibodies (Rheumatoid factor (RF) and anti-cyclic citrullinated peptide (CCP)) have proven very useful in predicting which patients will develop severe disease.

RA is the most common form of inflammatory arthritis, with an estimated worldwide prevalence of up to ~1% in the adult population. Females are at greater risk than males for developing the disease, with a female:male ratio of ~2.5:1. While the disease can occur at any age, the peak age on onset is in the 40’s, with an increasing incidence with age. To date, only two genes have been convincingly demonstrated to influence risk of RA (*HLA-DRB1* and the tyrosine phosphatase, *PTPN22*).

It has been estimated that the MHC region of the human genome accounts for approximately one-third of the overall genetic component of RA risk [1,2]. Genome-wide linkage scans using both microsatellite [3-7] and single nucleotide polymorphism (SNP) markers [8] have consistently identified this region as important in RA pathogenesis. These genome-wide scans have demonstrated that the MHC confers the largest genetic contribution in RA, and the relative contribution of HLA genes (*λHLA*) was found to be ~1.75 [3,5].
Much, but probably not all of the risk attributable to the MHC region is associated with alleles within the \textit{HLA-DRB1} gene. An association between RA and the class II HLA proteins was first noted in the 1970’s, when the mixed lymphocyte culture (MLC) type \textit{Dw4} (related to the serological subtype \textit{DR4}) was observed to be more common among patients with RA compared to controls [9,10]. Subsequently, investigation of the molecular diversity of class II proteins (subunits of \textit{HLA-DR, -DQ and -DP}) localized the serological \textit{Dw4} subtype to the \textit{HLA-DRB1} gene [11,12]. When the susceptible \textit{DR} subtypes were considered as a group, Gregersen et al noted a shared amino acid sequence at positions 70-74 of the HLA-DRB1 protein [13]. These alleles are now known collectively as “shared epitope” alleles due to the related sequence composition in the third hypervariable region: the susceptibility alleles result in missense amino acid changes, where the shared susceptibility amino acid motif is $^{70}Q/R-K/R-R-A-A^{74}$. Since the initial observation, a large number of population studies have confirmed the association between RA and allelic variants at \textit{HLA-DRB1} [14]. At the level of DNA, the most common (>5% population frequency) \textit{HLA-DRB1} shared epitope susceptibility alleles include *0101, *0401, and *0404 in individuals of European ancestry, and *0405 and *0901 in individuals of Asian ancestry; less common shared epitope alleles include *0102, *0104, *0408, *0413, *0416, *1001, and *1402. Of note, the *0901 allele observed among Asian populations does not strictly conform to the SE amino acid sequence motif ($^{70}R-R-R-A-E^{74}$), and the classic SE alleles may not contribute to risk in African-American and Hispanic-American RA populations [15,16].
While *HLA-DRB1* susceptibility alleles are often considered as a group, the strength of the genetic association to RA susceptibility differs across the *DRB1* alleles. There are at least two classes of *HLA-DRB1* risk alleles (high and moderate). In general, the *DRB1*\(^*0401\) allele exhibits a high level of risk, with a relative risk (RR) of approximately 3. The *DRB1*\(^*0101\), *\(^*0404\), *\(^*1001\), and *\(^*0901\) alleles exhibit a more moderate relative risk in the range of 1.5. These relative risk estimates are for all cases of RA. However, it is becoming increasingly clear that *HLA-DRB1* shared epitope alleles only influence the development of seropositive RA, and more specifically anti-CCP+ RA [17,18]. Collectively, the shared epitope alleles have an odds ratio of over 5 if CCP+ RA patients are compared to matched healthy controls. Because these alleles are quite common in the general population (collectively, allele frequency ~40% in individuals of European ancestry), the attributable risk for SE alleles is quite high.

Several investigators have proposed a refined classification of shared epitope alleles, as this hypothesis alone cannot explain all of the genetic risk attributable to the *HLA-DRB1* locus [19-22]. No consensus has emerged, however. Some of these studies suggest that a protective allele may be in linkage disequilibrium with the *HLA-DRB1* alleles. It has been hypothesized that the presence of an asparagine amino acid at position 70 of the HLA-DRB1 protein (D70) may be associated with protection from the development of RA (once the effect of the shared epitope alleles has been taking into consideration) [23,24]. Several studies suggest that additional genes within the MHC likely contribute to disease susceptibility, once the effect of *HLA-DRB1* has been taken into
consideration [25-29]. For example, an extended haplotype that includes HLA-DRB1*03 alleles may be associated with RA [25]. The associated haplotype spans ~500kb, and contains MHC class III genes, including the TNF locus implicated in other studies [28,30,31]. One study suggests that this association is restricted to CCP-negative patients [18].

Genetic variation at the HLA-DRB1 gene is clearly associated with rheumatoid arthritis (Figure 2). Our results suggest the existence of at least two classes of risk alleles (high and moderate). The *0401 and *0405 alleles have odds ratios in the range of 3.5, with lower-limit confidence intervals >3.0. The HLA-B54 allele is in LD with DR4 alleles (0405 in particular), which likely explains the association. The next class of HLA-DRB1 alleles have OR in the 1.5-2.0 range. These alleles include: *0101, *0404, *0901, and *1001; additional DR4 alleles are also included in this group (Dw14_0404-0408 alleles).

The *0401 allele is the only HLA-DRB1 allele with amino acid sequence 70Q-K-R-A-A74 in the third hypervariable region; the *0401 and *0405 alleles differ at amino acid 71 (where *0405 has an R). The *0405 allele, however, has the identical sequence as the *0404 allele. Thus, if the OR we observe for *0404 and *0405 are indicative of the true OR for these alleles, then a simple molecular model that includes variation at these critical amino acids is insufficient to explain the risk conferred by these alleles. In addition, our results do not support the classification schemes proposed by Gao and Michou [19,22], which differentiate between risk conferred by the *0401 and *0405 alleles.
Our results also support the hypothesis that genetic variation located a significant genomic distance away from the *HLA-DRB1* gene – and thus possibly not in LD with any of the *HLA-DRB1* risk alleles – also appears associated with RA risk. For example, alleles near the *TNF* locus appear associated with RA (e.g., *TNFa13*, OR 1.5).

REFERENCES


