GAW16 Data from the North American Rheumatoid Arthritis Consortium (NARAC)

The GAW16 RA data is the initial batch of whole genome association data for the North American Rheumatoid Arthritis Consortium (NARAC) cases (N=868) and controls (N=1194) after removing duplicated and contaminated samples. Total sample N=2062.

Background

The HLA region on 6p21 has been implicated by numerous studies, and there is consistent evidence that DR alleles contribute to disease risk. The 'shared epitope' hypothesis was proposed by Gregersen et al. (1987) to explain the organization of risk for rheumatoid arthritis from DR alleles. According to this hypothesis, individuals who share an QK/RRAA motif in positions 70-74 of the DR molecule show an increased risk for disease. The alleles that confer increased risk for rheumatoid arthritis include DRB1*0101, 0102, 0104, 0105, **0401, 0404, 0405, 0408, 0409,** 1001, 1402 and 1406, with highest risk alleles being bolded (Newton et al., 2004). This model was not quite sufficient to explain risk according to DR types and a newer model utilizing data from positions 70-74 has been developed (du Montcel et al., 2005). Aside from these main effects, there is also evidence for an interaction or haplotypic effects including the class 1 region and the central MHC, along with certain DR alleles, notably DR3 (Jawaheer, 2003,Irigoyen 2005).

Specific autoantibodies are noted to co-occur with rheumatoid arthritis. Rheumatoid factor IgM is a measure of active disease correlated with erosive arthritic disease. However, a more newly identified autoantibody, anti-cyclic citrullinated peptide (anti-CCP), is more specific for the disease and is a better predictor of erosive outcome (Huizanga et al., 2005). Elevations of anti-CCP have been noted to predict increased risk for development of rheumatoid arthritis (Kroot et al., 2000). The shared epitope alleles are strongly associated with the presence of anti-CCP antibodies, and this effect is modulated by HLA-DR3 (Irigoyen et al 2005). Alleles at the PTPN22 locus have been shown to confer an increased risk for RA (Begovich et al., 2004). At least two alleles of PTPN22 have been implicated as causing increased risk for RA; with the R620W allele in rs2476601 (hCV16021387) conferring 1.7-1.9 fold increased risk to heterozygotes and higher risks to homozygous carriers. Increased risk was also noted for either hCV8689108 or hCV25762283 (Carlton et al., 2005), with some indeterminacy because of LD among these markers (and others in the region).

Quantitative Phenotypes: Two quantitative phenotypes that are used for identifying RA affected individuals have been measured include anti-CCP and Rheumatoid Factor IgM. The heritability of these measures is hard to obtain from the selected sib pairs we are studying. After proband correction the heritability estimates are 11% and 30% while before correction the heritabilities are 15% and 67%. Genetic linkage analyses of these quantitative traits are underway and

results show major differences in chromosomal regions showing evidence for linkage between the two factors.

The Problem 1 Data

The NARAC data are contained in two files:

narac.csv

A comma-delimited file containing a header line and 2062 records. As indicated in the header line, each record contains the following fields:

ID	subject ID
Affection	RA affection status (0 = unaffected/control, 1 = affected/case)
Sex	gender (F = female, M = male)
DRB1_1	HLA-DRB1 allele 1
DRB1_2	HLA-DRB1 allele 2
SENum	indicates number of shared-epitope alleles (NN = 0, SN = 1, SS = 2)
SEStatus	shared-epitope alleles? (yes or no)
AntiCCP	anti-CCP
RFUW	rheumatoid factor IgM from University of Washington

545080 SNP-genotype fields from the Illumina 550K chip.

Missing values for CCP and RFUW are coded as question marks (?).

The genotypes are in the format X_X, where X is a base (A,C,G,T). Missing genotypes are coded as ?_?.

narac.map

A space-delimited file containing a header line and 545080 records (one per SNP). The SNPs are in the same order as they occur in the data file. As indicated in the header line, each record contains the following fields:

SNP	SNP name
Chromo	chromosome (0 = mitochondrial, 23 = X, 24 = Y, 25 = pseudo-autosomal)
Position	SNP position in basepairs

The number of fields in the comma-delimited data file may present a problem for some applications, e.g. Excel, and for Unix utilities, e.g. awk. We have written a simple Perl script, to be included on the CD, which will extract selected fields from the data file. The syntax for calling the script is

getdata -p -f *from* -t *to* -c *chr#* -o *output-file*

The selected fields are returned in comma-delimited (.csv) format. Subject IDs are always included. Command options are:

-p extract the phenotype fields (alone or in combination with genotypes)

-f extract SNP genotype fields beginning with ***from*** where ***from*** is either a SNP name or a basepair location

-t extract SNP genotype fields up to and including ***to*** where ***to*** is either a SNP name or a basepair location

-c chromosome number (required when the -f and -t options specifiy basepair locations, ignored otherwise)

-o name for the output file

Relevant Papers:

Begovich AB, Carlton VEH, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, Ardlie KG, Huang Q, Smith AM, Spoerke JM, Conn MT, Chang M, Chang S-YP, Saiki RK, Catanese JJ, Leong DU, Garcia VE, McAllister LB, Jeffery DA, Lee AT, Batliwalla F, Remmers E, Criswell LA, Seldin MF, Kastner DL, Amos CI, Sninsky JJ, Gregersen PK. A missense SNP in the protein tyrosine phosphatase *PTPN22* is associated with rheumatoid arthritis. *Am J Hum Genet* 2004;75:330-337.

Carlton VE, Hu X, Chokkalingam AP, Schrodi SJ, Brandon R, Alexander HC, Chang M, Catanese JJ, Leong DU, Ardlie KG, Kastner DL, Seldin MF, Criswell LA, Gregersen PK, Beasley E, Thomson G, Amos CI, Begovich AB. PTPN22 genetic variation: evidence for multiple variants associated with rheumatoid arthritis. *Am J Hum Genet* 2005;77:567-581.

du Montcel, S.T. et al. New classification of HLA-DRB1 alleles supports the shared epitope hypothesis of rheumatoid arthritis susceptibility. *Arthritis Rheum* 2005;52:1063-1068.

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.

Huizinga TW, Amos CI, van der Helm-van Mil AH, Chen W, van Gaalen FA, Jawaheer D, Schreuder GM, Wener M, Breedveld FC, Ahmad N, Lum RF, de Vries RR, Gregersen PK, Toes RE, Criswell LA. Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins. *Arthritis Rheum* 52:3433-3438, 2005.

Irigoyen P, Lee AT, Wener MH, Li W, Kern M, Batliwalla F, Lum RF, Massarotti E, Weisman M, Bombardier C, Remmers EF, Kastner DL, Seldin MF, Criswell LA, Gregersen PK. Regulation of anti-cyclic citrullinated peptide antibodies in

rheumatoid arthritis: Contrasting effects of HLA-DR3 and the shared epitope alleles. *Arthritis Rheum* 2005;52:3813-3818.

Jawaheer D. et al. Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families. *Arthritis Rheum* 2003;48, 906-916.

Kroot EJ, de Jong BA, van Leeuwen MA, Swinkels H, van den Hoogen FH, van't Hof M, et al. The prognostic value of anti-cyclic citrullinated peptide antibody in patients with recent-onset rheumatoid arthritis. *Arthritis Rheum* 2000;43:1831-1835.

Newton JL, Harney SM, Wordsworth BP, Brown MA. A review of the MHC genetics of rheumatoid arthritis. *Genes Immun* 2004;5:151-157.