GAW 15 Problem 3: Simulated Rheumatoid Arthritis Data

Full Model and Simulation Parameters

Michael B Miller <mbmiller@taxa.epi.umn.edu>,
Michael Li <nali@umn.edu>,
Gregg Lind <lind1199@umn.edu>,
Soon-Young Jang <jangsoon@taxa.epi.umn.edu>

School of Public Health, University of Minnesota

See the “readme” files distributed with the Problem 3 data for basic information about the data. This document explains the genetic model used to generate the data. For a quick overview of the model, look at the figure on the final page of this document and the figure caption on the penultimate page. If you have any questions after reading this document, please write to Mike Miller at the email address given above.

This document includes sections on the trait loci and their relationships to the phenotypes, more description of how the phenotypes were modeled, then details on the selection and modeling of marker loci.

Trait Loci

The following two tables summarize the trait locus positions and effects for all nine major genes. More details follow the tables.

Effects of Trait Loci (Major Genes)

locus	chr	cM	trait locus effect
DR	6	49.45557055	affects risk of RA
A	16	26.28791825	controls effect of DR on RA risk
B	8	170.9086911	controls effect of smoking on RA risk
C	6	49.45557055	0 cM from DR, increases RA risk only in women
D	6	54.57172282	5.12 cM from DR, rare allele increases RA risk 5-fold
E	18	94.27287711	controls effect of DR on anti-CCP and increases RA risk
F	11	115.286431	QTL for IgM
G	9	49.39545246	2 cM from Locus H, is 25% QTL for severity
H	9	51.41340983	2 cM from Locus G, is 25% QTL for severity

Genetic and Physical Map Locations of Trait Loci (Major Genes)

locus	chr	cM	Male cM	Female cM	physical location (bp)
DR	6	49.45557055	44.6111411	54.30000000	32484648
A	16	26.28791825	35.34662768	17.22920882	12517558
B	8	170.9086911	122.5730624	219.2443199	143800709
C	6	49.45557055	44.6111411	54.30000000	32484648
D	6	54.57172282	46.13944421	63.00400143	37233784
E	18	94.27287711	71.78986111	116.7558931	66069838
F	11	115.286431	73.75204226	156.8208197	110235253
G	9	49.39545246	47.61051259	51.18039234	26259317
H	9	51.41340983	48.05473497	54.77208468	27537313

We use “DR” as shorthand for the DRB1 locus of HLA. All loci are in linkage equilibrium with all other loci except for locin on chromosome 6. Loci C and DR have the maximum possible LD between them (a D' of 1.0), given their frequencies. Locus D is in very weak LD with Loci DR and C. All loci are diallelic except for DR which is triallelic.

Effects of Loci and Risk Variables Affecting Outcomes

Hazard and Risk

The model uses a constant hazard function to determine risk of RA. We planned at first to determine age at onset according to this exponential survival model, but it turned out that age of onset was then too strongly linked to some loci. We then retained the hazard approach but gave every individual the same risk period and the same “base hazard” (exponentiated intercept term). Therefore, multiplying hazard by some value is equivalent to multiplying risk by that value and the terms “risk” and “hazard” are used somewhat interchangeably below. Once hazard was known for a subject, we used the hazard to determine the mean of an exponential random variable. If this variable was less than a fixed threshold value (i.e., within the risk period), the subject was affected. The values of the base hazard and threshold are arbitrary, but they jointly determine population prevalence. Also see the section “RA affection status” under “Phenotypes” below.

HLA and Locus C

We model three DR alleles at the HLA locus. DR is in strong LD (multi-allelic D' = 1.0) and complete linkage (0 recombination fraction) with Locus C. DR effects are independent of locus C effects, but DR effects are epistatically controlled by Locus A. In females only, each C allele increases risk by a factor of 2.1 (female risk of RA is multiplied by 2.1 for the Cc genotype and by 4.41 for the CC genotype). Females with

no C alleles (cc) have no increased risk. The allele frequency for C is 0.5. DR/C haplotypes are shown in the following table.

DR/C Haplotype Frequencies (showing LD)

	C	c	
DR4	0.2500	0.0000	0.25
DR1	0.1000	0.0000	0.1
DRx	0.1500	0.5000	0.65
	0.5000	0.5000	1

Multiallelic: D = .15, D' = 1.0.

HLA/DR and Locus A

Locus A affects the impact of HLA DR types in a dominant fashion. Individuals with Aa or AA genotypes have their hazard multiplied by a value that is determined by their DR type according to the “Risk Multipliers” table below. A value of 1 indicates no change in risk. The allele frequency for A is .3 (thus the "aa" genotype has frequency 49% and "A_" has frequency 51%).

Average DR Risk (across A Genotypes)

	DRX	DR1	DR4
DRX	1	1	5
DR1	1	1.5	6
DR4	5	6	30

DR Risk Multipliers

	DRX	DR1	DR4
DRX	0.8	1	1
DR1	1	6	6
DR4	1	6	2

DR Risk ("aa") - 49% frequency

	DRX	DR1	DR4
DRX	1.11359	1	5
DR1	1	0.42254	1.69014
DR4	5	1.69014	19.86755

HLA Risk ("Aa" or "AA") - 51% frequency

	DRX	DR1	DR4
DRX	0.89087	1	5
DR1	1	2.53521	10.14085
DR4	5	10.14085	39.7351

Locus B

In smokers only, Bb or BB genotype multiplies RA risk by 1.5. This has the effect that smokers have no directly increased risk if their genotype is bb, but they still have some indirectly increased risk through the effect of smoking on IgM. The allele frequency for allele B is .35.

Locus D

Locus D has a direct effect on RA risk but a low allele frequency. Each D allele multiplies hazard by 5. The D allele frequency is only .0083 (exactly 1/120; so DD homozygotes are very rare).

Locus E

This locus has a strong direct effect on RA hazard, multiplying by 2.2 for each E allele (2.2 for Ee and 4.84 for EE). Locus E also affects anti-CCP by controlling which DR genotypes place a subject in the "high-mean" anti-CCP group (see the Anti-CCP section below). For a DR4 homozygotes only, having any E alleles puts one in the high-mean group. The high-mean group consists entirely of DR4 homozygotes with Ee or EE genotypes. The frequency of the E allele is .25.

No E Alleles (ee)

	DRX	DR1	DR4
DRX	-	-	-
DR1	-	-	-
DR4	-	-	-

One or Two E Alleles (Ee or EE)

	DRX	DR1	DR4
DRX	-	-	-
DR1	-	-	-
DR4	-	-	+

The “+” in the tables above denotes being in the group with high mean anti-CCP, which occurs only for DR4/DR4 homozygotes with at least one Ee or EE genotype.

Locus F

An additive effect of locus F causes 30% of the variance in IgM. Mean values of IgM are proportional to number of F alleles. The frequency of the F allele is .5.

Loci G & H

These two diallelic loci have allele frequencies of .1 and .2, respectively, and each contributes an additive genetic effect that accounts for 25% of the variance of latent severity (a total of 50% jointly). These loci are 2 cM apart on chromosome 9, but they are not in LD. Thresholds on latent severity are used to produce observed severity.

Phenotypes

Age. Ages for pairs of siblings were drawn from a bivariate normal distribution having parameters similar to pairs of affected siblings in real RA data we were given (rho= .855, stdev = 11.51, mean = 54.60), but pairs were retained only if both ages were between 18 and 87. The mother's age at the birth of the oldest sibling was uniformly distributed between 20 and 30 years and the father's age was equal to the mother's age plus a triangular random variable with a range from -1 to 5 and a mean of 2. This kept all ages reasonable and within acceptable ranges. The age reported for deceased individuals is the age they would have been at ascertainment of their oldest child, if they had lived. Age at death is also reported for deceased parents.

Sex. The sex of offspring was determined by age from published census data on sex ratio by age.

Death and Age of Death. All offspring are living. The variable "dead" has value 1 for parents who were deceased at the time of ascertainment, and value 0 for parents who were living. A parent was determined to be dead based on 2002 CDC mortality statistics for 10-year age classes. We applied a constant hazard within all but the oldest age group and started at the age of the parent when the youngest child was born. In the oldest age group from 85 to 100 years, the density for age at death had a linear form with the mode at 85 and zero density at 100, limiting longevity to age 100. We present data for dead parents as if they were alive. Age at death is provided. RA has a small mortality effect. The age of death for affected parents is on average 2 years (symmetrical triangular distribution with endpoints 0, 4) earlier than expected.

RA affection status. Affection was determined by taking a fixed threshold on an exponential random variable (values below threshold were affected). The mean of the exponential random variable (reciprocal of the hazard) was determined by multiplication of risk factors. More precisely, the log-hazard was modeled as a linear function of risk

factors and the individual exponential mean was 1/exp(log-hazard). This is a proportional hazards model with constant hazard and fixed follow-up time. Mortality and age were ignored in determining affection status. Variables and parameters that determined hazard are described below.

Smoking status. This was based on an age-dependent threshold model. A normal (0,1) random variable was generated for every subject such that variance was due to additive polygenic (50%), shared environmental (40%) and non-shared environmental (10%) influences. These numbers were based on results of a published twin study. Parents were genetically independent. Thresholds were determined by age according to CDC data so that individuals whose normal value exceeded a threshold were considered to be lifetime smokers at a probability appropriate for their age.

IgM. We generated a latent IgM value from a normal mixture with means determined by Locus F. Variance in latent IgM is caused by Smoking Status (24%), additive effect of Locus F (30%), and a residual (46%) with the residual variance being divided between additive polygenic (60%), and non-shared environmental (40%) components. The IgM latent variable was transformed monotonically to fit the distribution of IgM in real RA data.

Anti-CCP. Locus E and HLA/DR genotype jointly created 10.3% of the variance in anti-CCP as described in the "Locus E" section above. The remaining variance was caused by additive polygenic (60%), and non-shared environmental (40%) components. The anti-CCP latent variable was rank rescaled using the observed distribution of values in the RA reference data to create the final anti-CCP values.

Severity. Severity was determined by two diallelic loci (G and H, allele frequencies of 0.1 and 0.2 respectively) with additive effects. Each of the loci accounts for 25% of the total variance. The remaining variance (50%) is due to an individual random environment effect. There are 5 severity classes, each containing 20% of the affected persons.

Age of Onset. The age of onset (for affected offspring) was created from an "onset" latent variable that equally weighed the hazard, latent severity, and an independent random variate. This variable was converted to ranks and used with real RA data to derive a "proportion of life affected," which multiplied by the ascertainment age, yielded the age of onset.

Residual Effect: There is a residual effect on the log-hazard for RA that is composed of shared environment effect (85% of variance) and a non-shared environment effect (15%). The shared environment effect is a constant multiple of a Bernoulli-distributed random variable and it is shared by all members of a family in 30% of families. The non-shared environment effect was normally distributed. The convolution (sum of the two variables) was a normal mixture with a standard deviation of 2.079.

Summary Of Key Covariate Effects

Smoking: Affects RA risk directly with Locus B interaction, and through its effect on IgM.

Age: Age affects RA risk through its affect on smoking and through its effect on the sex ratio. Age affects mortality, but only in the parents, and we report affection status regardless of mortality.

Sex: Nearly all of the sex effect comes from Locus C, but the general population M:F sex ratio in the offspring generation has an effect.

Use of HapMap Data

All trait loci and marker data were derived from HapMap Phase I data on 120 haplotypes estimated from CEPH data on Utah Residents with Northern and Western European ancestry (known as the "CEU" data). This naturally created LD between all markers and trait loci.

Generating Founder Haplotypes

The haplotypes for founders in our nuclear families were generated by randomly recombining HapMap CEU haplotypes. Based partly on population genetics theory and partly on our tests of LD, we chose to multiply map length by 30 so that on a 1 Morgan interval of a founder haplotype there are, on average, 30 points where the ancestry switches from one of the 120 HapMap haplotypes to another one (but the "other one" would be the same one with probability 1/120). The number of these switch points per Morgan follows a Poisson distribution and locations of switch points are uniformly distributed in the genetic map. This provides something very similar to what would be obtained by selecting a random haplotype after 30 generations of random mating of a large population of the 120 HapMap haplotypes where the HapMap haplotypes have equal frequency. (It differs only in using a sex-averaged map instead of allowing for male-female differences in recombination -- but the effect of modeling such an effect would be trivial. We used sex-specific maps to model genetic transmission within families.)

Generating maps

HapMap data provided estimated sex-averaged map locations in cMs for every SNP, but it did not provide sex-specific maps. To derive sex-specific maps for the HapMap markers, we used the sex-specific maps presented by Kong et al.[1] and we used linear interpolation to choose appropriate male and female distances that corresponded to the

sex-averaged distances from HapMap. The sex-averaged maps were used to generate the locations of recombination junctions for meioses in the nuclear families.

Recombination

We used a Haldane model (Poisson process). The locations of all recombination junctions from transmitted chromosomes were stored and used for every marker set (trait loci, microsatellites, 10k SNPs and Chromosome 6 dense SNP map). We also store complete information about how parental haplotypes were extracted from HapMap data. Thus, our method allows us to add more markers from the HapMap data after the simulation project is over, just as one could type more markers from DNA samples.

Selection of Microsatellites

To generate microsatellites, we decided to look at every set of four consecutive SNP markers in hapmap and treat them as binary numbers from 0 to 15. We could then treat the binary numbers as alleles. Each marker so constructed was placed on the map at the location of the left-most of the four SNPs used to construct it. We then generated a list of the locations of the markers with heterozygosities exceeding .70. We chose how many of those polymorphic markers to retain based on the numbers of markers that had been used in a 5 cM Marshfield marker set. We then selected the subset of markers on every chromosome to achieve the target number of markers while maximizing the size of the smallest inter-marker distance and retaining the two most telomeric markers.

Microsatellites selected in this way would have a stronger LD with nearby SNPs than one would expect to see in real data. Therefore, we "mutated" the microsatellites by adding a random number. We added floor($3*U^4$), where U is a uniform (0,1) pseudorandom value. This term has a distribution of 0, 1, 2 with probabilities, .76, .14, .10. This weakens LD between microsatellites and nearby SNPs to some degree.

There are 730 microsatellite markers on the 22 autosomes with a maximum inter-marker distance of 9.3 cM (sex averaged).

Selection of SNP Markers for the 10K SNP Set

Our goal here was to mimic an Affymetrix Xba 131 10K SNP chip. The distribution of the SNP minor allele frequencies in HapMap is similar to that of Affymetrix 10K SNP chip, but we did not use any monomorphic SNPs. We simply ignored frequencies and focused on physical map positions. We took a list of the physical map locations of the SNPs from the Affymetrix 10k chip and for every one of those markers, we identified the marker in HapMap that was physically closest to the marker from Affymetrix. We then used that collection of HapMap markers for our 10K SNP set.

There are 9,187 SNPs on the 22 autosomes in the 10K SNP set and a maximum inter-marker distance of 11.8 cM.

Selection of a Dense SNP Map for Chromosome 6

We wanted the chromosome 6 SNP map to correspond roughly to what one might have for chromosome 6 from a 300K SNP chip. We counted the number of HapMap SNPs and found that there were about 812,000 SNPs, roughly three times the number on a 300K SNP chip. We then simply retained every third SNP from the HapMap data to produce 8,910 SNPs for the Chromosome 6 dense SNP map.

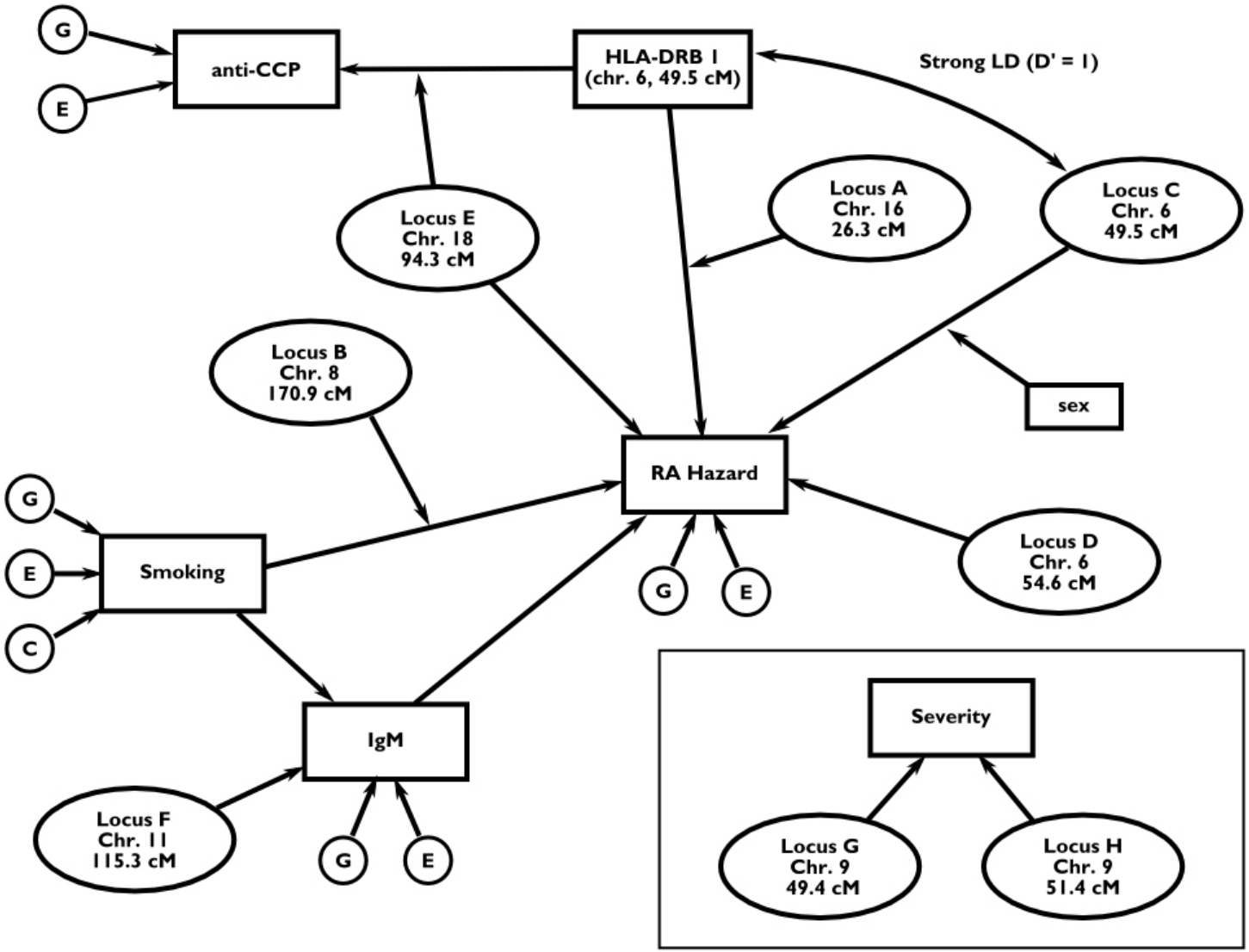
Acknowledgement

We thank the members of the GAW15 committee for their help and patience. We especially thank Chris Amos for coming up with most of the interesting ideas on how to model RA genetics, but please blame us for anything you don't like!

References

1. Kong A, et al. (2002). A high-resolution recombination map of the human genome. Nature Genetics, 31, 241-247.

Figure Caption. Genetic loci are represented as ovals, normally-distributed polygenic/environmental variables are represented as circles (G = additive polygenic, C = common family environment, E = non-shared environment) and observed variables are represented as rectangles. The RA hazard is a continuous variable that is dichotomized into affected/unaffected before it is observed, and severity is polytomized into five levels before it is observed. Arrows indicate where effects of variables are manifested. For example HLA-DRB1 affects both anti-CCP levels and RA Hazard, but the strength of its effect on anti-CCP is controlled by Locus E genotype and the effect of HLA-DRB1 on RA Hazard is controlled by Locus A genotype.


G
E
anti-CCP
HLA-DRB 1
(chr. 6, 49.5 cM)
Strong LD (D' = 1)
Locus E
Chr. 18
94.3 cM
Locus A
Chr. 16
26.3 cM
Locus C
Chr. 6
49.5 cM
Locus B
Chr. 8
170.9 cM
sex
RA Hazard
G
E
C
Smoking
G
E
Locus D
Chr. 6
54.6 cM
IgM
Locus F
Chr. 11
115.3 cM
G
E
Severity
Locus G
Chr. 9
49.4 cM
Locus H
Chr. 9
51.4 cM