# Problem 3: Simulated Rheumatoid Arthritis Data

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>

The plan for this data simulation was to mimic the familial pattern of rheumatoid arthritis (RA) including a strong effect of DR type at the HLA locus on chr 6 and other genetic and environmental effects. For each of 100 replicates, we generated a large population of about two million nuclear families (two parents and two offspring) with RA affection status determined by a complex genetic/environmental model, then we retained a random sample of 1,500 families from those families that had an affected sibling pair (ASP) and a random sample of 2,000 families where none of the four members were affected (control). We present data on all members of the 1,500 ASP families and on one randomly-selected member of the offspring generation from each of the 2,000 unaffected control families (i.e., there are 2,000 unrelated control subjects per replicate, and no control subject had a first-degree relative with RA).

Here are some results of analysis of data from a simulated general population of 1,800,000 sibling pairs (3,600,000 subjects) generated using the model we developed:

| | |
|---|---|
| RA lifetime prevalence | 0.0107 |
| F:M sex ratio in affecteds | 3.07 |
| Lambda_sib | 9.03 |
| Number of ASPs | 1856 (of 1.8 million sib pairs) |

"Lambda_sib" is the lifetime prevalence in siblings of affected individuals (probandwise concordance) divided by the lifetime prevalence in our simulated general population. The numbers in the table above are similar to what one would see in real RA epidemiological data. These numbers apply to the offspring generation, and not the parent generation of our simulated data.

We present markers on 22 autosomes which were designed to be like real human autosomes in terms of genetic and physical map lengths, but we did not generate data for sex chromosomes. The markers are in three sets:

1. A set of 730 microsatellite markers, fairly evenly spaced on chromosomes with an average inter-marker distance of about 5 cM and with heterozygosities always exceeding .70.
2. A set of 9,187 SNPs distributed on the genome to mimic a 10K SNP chip set but without monomorphic SNPs.
3. A very dense map of 17,820 SNPs on chromosome 6 (an average inter-marker spacing of 9,586 bp which corresponds roughly to the density one would expect from a genome-wide 300K SNP set). The chr 6 dense map includes 210 of the markers from the 10K SNP map (they are easily identifiable because they have the same names in both sets).

The data for marker and trait loci were generated so that they will have many of the properties of real data. For example, you should observe LD over short genetic distances both between pairs of markers and between markers and trait loci and the patterns of LD should be similar to patterns seen in real data.

## Data Files - Filenames and Formats:

The data consist of both map information, with lists of markers and their locations, and simulated marker and phenotype data. The map data are stored in maps.tar.gz, and they are described below following the discussion of the simulated data which are stored in a different collection of files.

### *Simulated Marker and Phenotype Data Files*

The marker and phenotype data from the 100 replicates are stored in gzipped tar files named rep0001.tar.gz, rep0002.tar.gz, ..., rep0100.tar.gz. Each of these compressed files uses about 79 MB of disk space, but when uncompressed and extracted they create directories named rep0001, rep0002, ..., rep0100, each of which uses about 891 MB of disk space. To uncompress and extract the files from the compressed archives, the following command seems to work on any unix system that has gunzip installed:

```
gunzip -c rep0001.tar.gz | tar xvf -
```

Simply replace *0001* with appropriate digits to extract other replicates. If you are using Linux, Cygwin, FreeBSD, Mac OS X, or some other Unix system, you can use this shorthand command to accomplish the same thing:

```
tar zxvf rep0001.tar.gz
```

If you are using Windows, we recommend that you either install Linux in a dual-boot configuration with Windows or run Cygwin under Windows (because of all the additional functionality those options provide for genetic analysis), but if those ideas are unappealing, the program 7-zip (http://www.7-zip.org/) should work for you. Unfortunately, 7-zip requires that uncompressing and extracting be done in two steps which takes up about 891 MB of extra disk space temporarily.

If you are using Mac OS 9, the data files can be extracted using Stuffit Expander (http://www.stuffit/mac/expander/).

The names of files within the replicate directories have the formats shown below where ***** stands for the replicate number padded with zeros (0001, 0002, ..., 0100), **#** stands for the chromosome number, **ASP** refers to the affected sibling pair nuclear families, **CONTROL** to the unrelated control subjects, **SNP** to SNP marker data, **STRP** to microsatellite data and **dense.SNP** to the dense SNP data from chromosome 6. The **PHENOTYPE** files contain all of the phenotype data for ASP families and control subjects.

*Phenotype data filenames*:

ASP.*.PHENOTYPE.ped
CONTROLS.*.PHENOTYPE.ped

*Marker data filenames*:

chr#.*.CONTROLS.SNP.ped
chr6.*.CONTROLS.dense.SNP.ped
chr#.*.ASP.SNP.ped
chr6.*.ASP.dense.SNP.ped
chr#.*.ASP.STRP.ped

Thus, for every replicate, there are 68 files of marker data (22 chromosomes times two types of subjects [ASP nuclear families and control subjects] for SNPs, 22 chromosomes of microsatellites for ASP families only and dense SNPs for chromosome 6 for both ASP families and controls) and two phenotype files: a total of 70 data files per replicate.

All fields (data columns) in every pedigree file (*.ped) are delimited by a single space. The first five columns always consist of these fields:

<Family ID> <Individual ID> <Father ID> <Mother ID> <Sex>

Family IDs are strings consisting of replicate number, underscore, family number (e.g., family ID **71_1367** corresponds to the 1,367th family in the 71st replicate). All individual IDs are unique integers within every replicate, but all replicates use the same sets of individual Ids. Control subject individual IDs (numbers 10001 to 12000) differ from ASP individual IDs (numbers 1 to 6000). This system makes it easy to mix together data from cases and controls from the same replicates or from different replicates. Following the first five columns are either phenotypes (in the *.PHENOTYPE.ped files) or markers. The markers always consist of two alleles per marker, separated by a single space (in other words, there are two space-delimited fields per marker). All alleles are integers from 1 to 20, but they are not necessarily consecutive integers because some alleles might not be observed. All SNP loci are diallelic and alleles are coded as **1** and **2**.

Consecutive columns of the phenotype data (*.PHENOTYPE.ped files) are separated by a single space. After the first five columns described above, the remaining eleven columns consist of phenotypes in the following order:

6 Rheumatoid arthritis affection status (2=affected, 1=unaffected)
7 Dead (1=dead, 0=not dead)
8 Age at ascertainment (in years)
9 Lifetime smoking (1=smoked, 0=never smoked)
10 Anti-CCP continuous measure
11 IgM continuous measure
12 Severity (1 to 5; 1=mild, 5=severe)
13 DR allele from father
14 DR allele from mother
15 Age at death (missing if alive)
16 Age at onset (only available in offspring)

*Map Data Files*

The map data are stored in the file maps.tar.gz which expands to a directory named maps. The files within the maps directory have names with the following forms where # represents the chromosome number:

chr#.SNP.map
chr#.STRP.map
chr6.dense.SNP.map

The SNP and STRP in the filenames mean that the map data are for SNPs or for microsatellites (Simple Tandem Repeat Polymorphisms). Only chromosome 6 has a very dense SNP map. Within the .map files we report sex-averaged, male and female map locations (in Haldane cM) for all markers. We also present the physical location in base pairs. All fields in map files are separated by a single space and this is the format of every line of every map file:

<chr. no.> <marker name> <cM sex-ave.> <cM male> <cM female> <base pairs>

The tables below show the numbers of markers per chromosome for the three kinds of marker files:

***microsatellite markers***

| *count* | *file* |
|---|---|
| 67 | chr1.STRP.map |
| 56 | chr2.STRP.map |
| 49 | chr3.STRP.map |
| 38 | chr4.STRP.map |
| 45 | chr5.STRP.map |
| 41 | chr6.STRP.map |
| 41 | chr7.STRP.map |
| 31 | chr8.STRP.map |
| 35 | chr9.STRP.map |
| 38 | chr10.STRP.map |
| 32 | chr11.STRP.map |
| 36 | chr12.STRP.map |
| 24 | chr13.STRP.map |
| 28 | chr14.STRP.map |
| 22 | chr15.STRP.map |
| 27 | chr16.STRP.map |
| 26 | chr17.STRP.map |
| 27 | chr18.STRP.map |
| 17 | chr19.STRP.map |
| 22 | chr20.STRP.map |
| 11 | chr21.STRP.map |
| 17 | chr22.STRP.map |
| 730 | *total* |

***SNP markers***

| *count* | *file* |
|---|---|
| 704 | chr1.SNP.map |
| 813 | chr2.SNP.map |
| 687 | chr3.SNP.map |
| 642 | chr4.SNP.map |
| 622 | chr5.SNP.map |
| 674 | chr6.SNP.map |
| 479 | chr7.SNP.map |
| 442 | chr8.SNP.map |
| 475 | chr9.SNP.map |
| 472 | chr10.SNP.map |
| 492 | chr11.SNP.map |
| 496 | chr12.SNP.map |
| 406 | chr13.SNP.map |
| 334 | chr14.SNP.map |
| 257 | chr15.SNP.map |
| 204 | chr16.SNP.map |
| 156 | chr17.SNP.map |
| 303 | chr18.SNP.map |
| 93 | chr19.SNP.map |
| 187 | chr20.SNP.map |
| 174 | chr21.SNP.map |
| 75 | chr22.SNP.map |
| 9187 | *total* |

***chr 6 dense SNP markers***

| *count* | *file* |
|---|---|
| 17820 | chr6.dense.SNP.map |

Remember that combining the two chromosome 6 maps gives us a total of 18,284 distinct SNP markers on chromosome 6.

In addition to the map files, we supply some files in a maps/MERLIN directory that use the MERLIN .dat format to specify the contents of the marker map and phenotype files.

## Unusual Features of the Simulated Data:

We decided to provide more information in the simulated data than one would ordinarily have in real data. This allows the analyst to do some interesting things.

*No Missing Data:* One usually expect to be missing marker information on some family members, especially those who died before the family was ascertained. We provide marker data on all family members. Researchers who would like their data to be more realistic can delete marker information from deceased individuals. By supplying data that would normally be missing, we provide more opportunity to test effects of missingness, etc.

*No Errors:* We did not model any errors in the data simulation. In real data there are typically some errors in genotyping and sometimes there are sample mixups. By not modeling any errors, we make it possible for the analyst to simulate his own errors and test the effect of genotyping error on other aspects of a genetic analysis. We also added no errors to phenotypes.

*Allele ordering in the output:* The allele inherited from the father is always presented on the left side within every genotype. This allows researchers to determine haplotypes for all subjects and to determine their parental origin. In real data, it is usually not possible to know haplotypes or their origin like this, but new methods have made molecular haplotyping possible and it is currently being used. So, in real data we can sometimes know haplotypes, but the parental origins of those haplotypes still must be inferred.

We hope you enjoy these data. We'll see you at the GAW meeting in November.

We thank the Minnesota Supercomputing Institute for their support for this simulation project. This work was supported, in part, by NIH grants 5RO1-HL09609-12, 1R01 AG021917-01A1, and by the University of Minnesota.