Technical Document by SBIL, 2007.

Simulation of 317K SNP data

Preliminary investigation of the feasibility of applying and evaluating the performance of the statistical machine learning methods to the analysis of large numbers of SNPs was investigated via simulation. Data on 223 individuals that were genotyped on the 317K Illumina HumanHap300 BeadChip as part of the New York City Cancer Control Project (NYCCCP; PI Peter Gregersen) form the base from which the simulated data were generated. To facilitate this investigation, a flexible simulation program was written that generates user defined sample size, number of SNPs, no missing data consistent with the observed missing data in the original genome scan, and affected or unaffected disease status under the null hypothesis (i.e., no associations in the genome) or under the alternative hypothesis (i.e., hard coded penetrance functions). Missing data is "filled in" completely at random and proportional to the allele frequencies in the original data.

The simulation of these data proceeds as follows. Consider a matrix with 223 rows corresponding to NYCCCP individuals and 317,503 columns corresponding to the 317,503 SNPs. The elements of this matrix are the individual genotypes. Partition the columns into "bins" of 500 consecutive SNPs. That is, 636 bins, where the last bin only has 3 SNPs. The simulated genome scan data for each individual is obtained by random draws (with replacement) from real data matrix of 223 individuals and 636 bins of 500 SNPs. Specifically, the simulated data for an individual is generated by randomly selecting the first bin (first column) from the 223 individuals (rows), randomly selecting with replacement the second bin from the 223 individuals, randomly selecting with replacement the third bin from the 223 individuals, and so on through all 636. Thus the data retains the basic patterns of linkage disequilibrium, missing data, and allele frequencies as that observed in the original genome scan data. The exception to this is only at the 635 breaks in the genome corresponding to the bin boundaries. The same random process of sampling bins with replacement is repeated for all individuals in the resulting simulated dataset. Modifications of this program will include additional data from the NYCCP as it becomes available and refinement of the bin boundaries to correspond to the chromosome and centromere.

The simulations presented as preliminary data for this application correspond to approximately 1000 cases and 1000 controls simulated under the alternative hypothesis described below and no missing data. Only autosomal loci are considered in the preliminary data. The various statistical machine learning methods were applied to sets of 100, 250, 500, 1000 and 2000 SNPs selected at random from the autosomal loci. The simulations reported assume that the disease risk is 100% explained by genetic factors and no interactions with environmental effects at this time. However, the program has the capability for both normal and multinomial random variables that can be factored into a penetrance function as either an independent environmental factor or part of one or more polymorphisms by environment interaction.

The seven SNP-by-SNP interaction models were explored in the simulation. Five of the seven models involve interactions. None of the interactions are motivated by an additive or multiplicative model. Rather, all interactions are generated by more complex Mendelian inheritance pattern. Thus, additive genetic model statistical tests are not the optimal statistical test. A total of 17 SNPs influence disease status. For each of the analytic methods the total number and which individual SNPs were selected to be retained were recorded.

<u>Model 1 – Five locus interaction among common alleles, nearly fully penetrant.</u>

The model assumes a five-locus interaction under a dominant genetic model for the minor (less frequent) allele at each locus. It assumes a minor allele frequency of 0.30 at each locus. The expected number of such genotype combinations is approximately 35 under complete independence. The penetrance function (probability of disease) is zero if the minor allele is not present at each locus and 0.90 if the minor allele is present at each locus. In equation form it is:

$$\text{Prob}\left(\text{Disease}\middle| G^{A}_{12or22} \wedge G^{B}_{12or22} \wedge G^{C}_{12or22} \wedge G^{D}_{12or22} \wedge G^{E}_{12or22}\right) = 0.90, \text{ zero otherwise.}$$

Here, G_{12} indicates the genotype is 12 (i.e., heterozygous). The five loci are noted as the superscript, and ^ denotes intersection. Note: $G_{12}=G_{21}$.

Model 2 – Three locus interaction among reasonably common alleles, fully penetrant

The second model assumes a three-locus interaction. The minor allele frequencies at the three loci are 0.25 for A, 0.20 for B, and 0.20 for C. The model is a fully penetrant model (i.e., probability of disease is one given the predisposing genetic factors). In table format we have:

		G^A_{11}				G^A_{12}				G^A_{22}		
		G^C				G^C				G^C		
		11	12	22		11	12	22		11	12	22
	11	0	0	0		0	0	0		0	1	1
G^B	12	0	0	0		0	0	0		1	1	1
	22	0	0	0		0	0	0		1	1	1

Here, the set of three columns under G^A_{11} form a sub-table that corresponds to the possible genotypes at the other two loci G^B and G^C. In the first sub-table, the penetrance function or probability of disease is zero. Thus, only in the last sub-table corresponding to the recessive model a locus A is there a nonzero penetrance function. Specifically, it is a fully penetrant model under a recessive model for locus A and dominant model for loci B and C.

Model 3 – Three locus interaction, common alleles, incomplete penetrance

The third model assumes a three-locus interaction. The minor allele frequencies at the three loci are 0.40 for A, 0.25 for B, and 0.25 for C. Model 3 penetrance functions can be summarized as:

Prob(disease | heterozygous for at least two loci and not G^A_{11})=0.5.

Prob(disease | homozygous for minor allele at two loci and not G^A_{11})=1.0.

		G^A_{11}				G^A_{12}				G^A_{22}		
		G^C				G^C				G^C		
		11	12	22		11	12	22		11	12	22
	11	0	0	0		0	0	0		0	0	0
G^B	12	0	0	0		0	0.5	0.5		0	0.5	1
	22	0	0	0		0	0.5	1		0	1	1

Model 4 – Two locus interaction, common alleles, incomplete penetrance, dominant model

The fourth model assumes a two locus interaction for common alleles under a dominant genetic model at each locus. The minor allele frequencies are 0.20 for locus A and 0.30 for locus B. The penetrance function is summarized in the table below.

		G^B		
		11	12	22
	11	0	0	0
G^A	12	0	0.75	0.75
	22	0	0.75	1.0

Model 5 – Two locus interaction under a dominant model for the major allele.

The fifth model assumes a two locus interaction under a dominant model for the **major** allele. The model is for a very common but low penetrant allele. The minor allele frequencies at these two loci are 0.25 and 0.25, respectively.

		G^B		
		11	12	22
	11	0.10	0.10	0
G^A	12	0.10	0.10	0
	22	0	0	0

Model 6 – Single locus dominant model, partial penetrance of an uncommon allele.

Model 6 assumes a partially penetrant dominant model at one locus with a minor allele of 0.10. The penetrance function is Prob(disease | G_{12} or G_{22}) = 0.5 and zero otherwise.

Model 7 – A single locus model under a recessive genetic model and a common allele.

Model 7 assumes a partially penetrant recessive model at one locus with a minor allele of 0.40. The penetrance function is Prob(disease | G_{22}) = 0.5 and zero otherwise.

As part of the scope of work for the proposed project we plan to develop additional models that represent a wide array of Mendelian and more complex nonlinear multilocus interactions to be included in the panel of multi-locus interactions that will be the targets for discovery using the novel and conventional analytic methods.

Simulation dataset parameters

The simulation data for each iteration will be generated using the simulation program and the NYCCP control data as described in Section C.4.1. Briefly, the simulated data for each person will be generated based on random draws with replacement from 636 bins of the genome and all available NYCCP controls of European American descent. Prior to analysis all SNPs inconsistent with HWE expectations (P<0.01) will be removed. In addition, SNPs with >20% missing data or differences in missing data proportions between cases and controls (P<0.05) will be removed. Causal SNPs will be selected based on closeness to the designated allele frequencies, consistency with HWE (P>0.20), less than 2% missing data and no evidence of differences between case and control missing data proportions (P>0.10).

It is the purpose of this study to examine higher order interactions not easily detected by conventional contingency table or log-linear models. The penetrance functions for the causal clusters of SNPs or informative features (i.e., individual polymorphism effects, 2, 3, and 5 polymorphism interactions) will be defined similarly to that in section C.4.1. Each simulation will contain 17 predisposing SNPs: two main effect only SNPs, 2 two-way interactions, 2 three-way interactions and 1 five-way interaction. One two-way and one three way interaction will contain a binary simulated environmental factor (e.g., smoking yes/no, diabetes yes/no). Each simulation dataset will contain 1000 cases and 1000 controls. A total of 1000 simulated datasets (iterations) will be generated and analyzed. The above 1000 iteration experiment will be completed four different times corresponding to four different sets of 17 SNPs under different main effects/interaction relationships. Three of the four simulation experiments will focus exclusively on interactions described by locus-specific Mendelian genetic models (i.e., combinations of dominant, additive and recessive models). The last simulation experiment will focus on interaction models not described by simple locus-specific Mendelian genetic models. For two-way interactions, numerous examples can be obtained (see Table 1 of Li and Reich 2000, in Appendix II). For example, consider the interaction table where the penetrance is either 0 or 1 depending on if the cell is on one of the diagonals or off-diagonals, respectively.

	aa	aA	AA
bb	0	1	0
bB	1	0	1
BB	0	1	0

Power of the simulation study

Assuming an additive genetic model, a type 1 error rate of 0.01 and depending on minor allele frequencies, 1000 cases and controls will have locus-specific power of 0.80 to detect odds ratios between 1.2 and 1.5. Assuming a type 1 error rate of 0.05, a 1000 iteration experiment has 0.80 statistical power to detect differences of 0.064 in the proportion of causal SNPs correctly identified by two different methods; similar proportional difference estimates for 0.70 and 0.90 levels of power are 0.057 and 0.074, respectively. Thus, the simulation study design has the power to detect meaningful differences in the ability of the various statistical machine learning methods to detect informative features. Given the combined computing resources of the two institutions, time trials indicated that 1000 is a very feasible number of iterations for the simulation study.