Web-based Supplemental Materials for

Power of data mining methods to detect genetic associations and interactions

by Annette M. Molinaro, Nicholas Carriero, Robert Bjornson, Patricia Hartge, Nathaniel Rothman, and Nilanjan Chatterjee.

1 Alternative Algorithms

1.1 Monte Carlo Logic Regression

Similar to the steps described for Random Forest, an initial, observed, data set was simulated with n_{cases} and $n_{controls}$ using one of the three models outlined in Section 3.1. Due to the need for binary covariates in Logic Regression, all simulated variables were transformed to p^* dummy variables before being input to the algorithm, where $p^* = 2 * p$. If the original SNP value = 0, then both dummy variables = 0; if the original SNP value = 1 then the first dummy variable = 1 and the second is 0; if the original SNP value = 2 then the first dummy variable is 0 and the second =1. A null distribution of B = 100; 000 data sets was generated by permuting the outcome labels of the first simulated data set B times, running MCLR on each and recording the variable importance ratios. Subsequently, MCLR was run on the initial data set with un-permuted class labels, as well as the $nsim - 1$ additionally generated data sets, and the corresponding variable importance ratios were recorded. A p-value for the p^*th dummy variable was assessed using Equation 1, replacing $VI^O_{p^*}$ by $VIR^O_{p^*}$, the variable importance ratio for the p^*th dummy variable as assessed using the observed (non-permuted) data set, and $VI^b_{p^*}$ by $VIR^b_{p^*}$, the variable importance ratio for the p^*th dummy variable as assessed using the $b = 1, \ldots, B$ sample with permuted labels. The reported p-value for each original SNP is the sum of its corresponding dummy variables p-values out of $nsim$ that fall below a specified cut-off, e.g., $\alpha_{cut-off} = 0.05/p^*$ divided by $nsim$.

1.2 MDR

With the same steps as the other algorithms, with MDR a p-value for the pth variable is assessed using Equation 1, replacing VI^O_p by $VICV^O_p$, the variable importance proportion for the pth variable as assessed using the observed (non-permuted) data set, and VI^b_p by VIR^b_p, the variable importance proportion for the pth variable as assessed using the $b = 1; , \ldots, B$ sample with permuted labels. The reported value for each variable is the ratio of p-values out of $nsim$ that fall below a specified cut-off, e.g., $\alpha_{cut-off} = 0.05/p$.

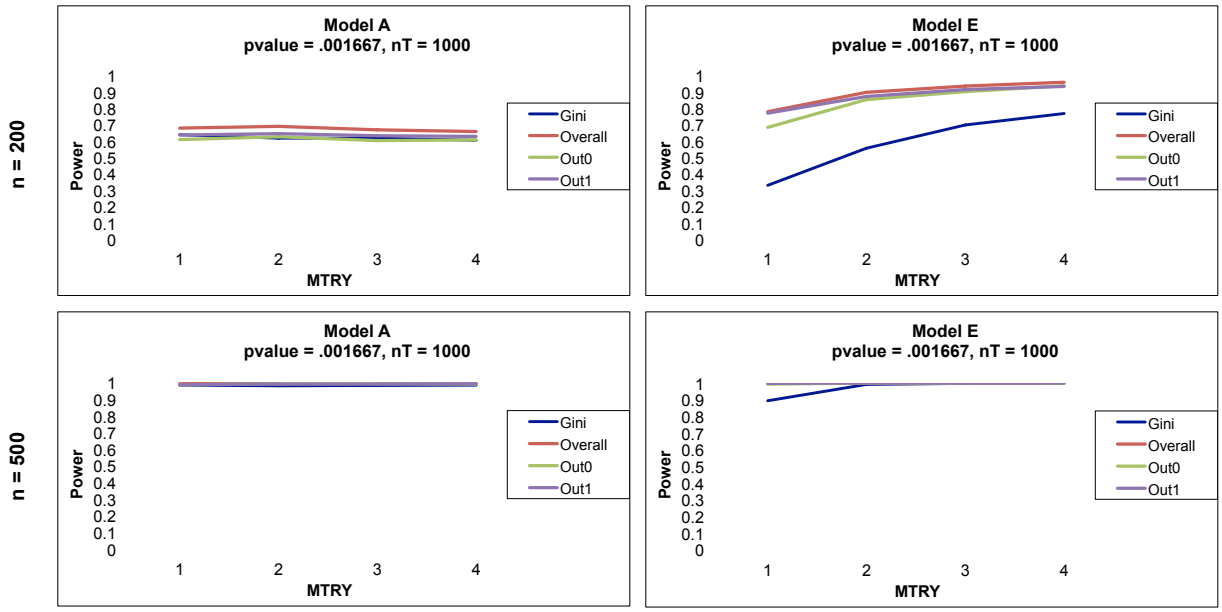

Figure 1: **Random Forests results for the Additive (Model A) and Exact (Model E) simulations.** Row 1 shows the results for 200 observations and row 2 for 500 observations with continuous variables. Power is measured on the y-axis and $mtry$ on the $x-$, p-value is adjusted for 30 SNPs, and $nT = 1000$. Colored lines correspond to RF' four variable importance measures.

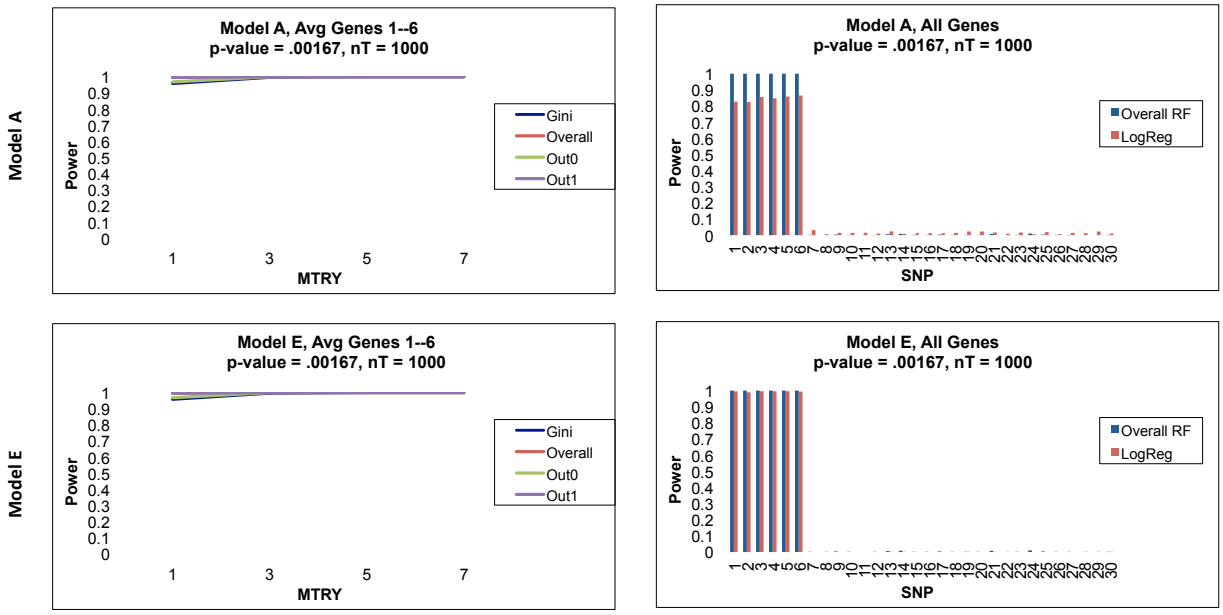

Figure 2: **Random Forest and Monte Carlo Logic Regression results for Additive (Model A) and Exact (Model E) simulations.** The results in the top row are for the Additive Model and the bottom row for the Exact Model. All displayed results are based on 500 observations per sample using dummy variables for the SNPs. For each, the power is measured on the y-axis. In the first column only RF results are illustrated with the four different values for mtry on the x-axis. The number of trees for each forest (nT) is equal to 1000. The achieved power, averaged over the first six SNPs, for each of the four variable importance measures are shown in colored lines. In the second column, the results for RF (based on the overall measure of variable importance) are compared to MCLR over each SNP (x-axis). The p-value is adjusted for the 30 SNPs.

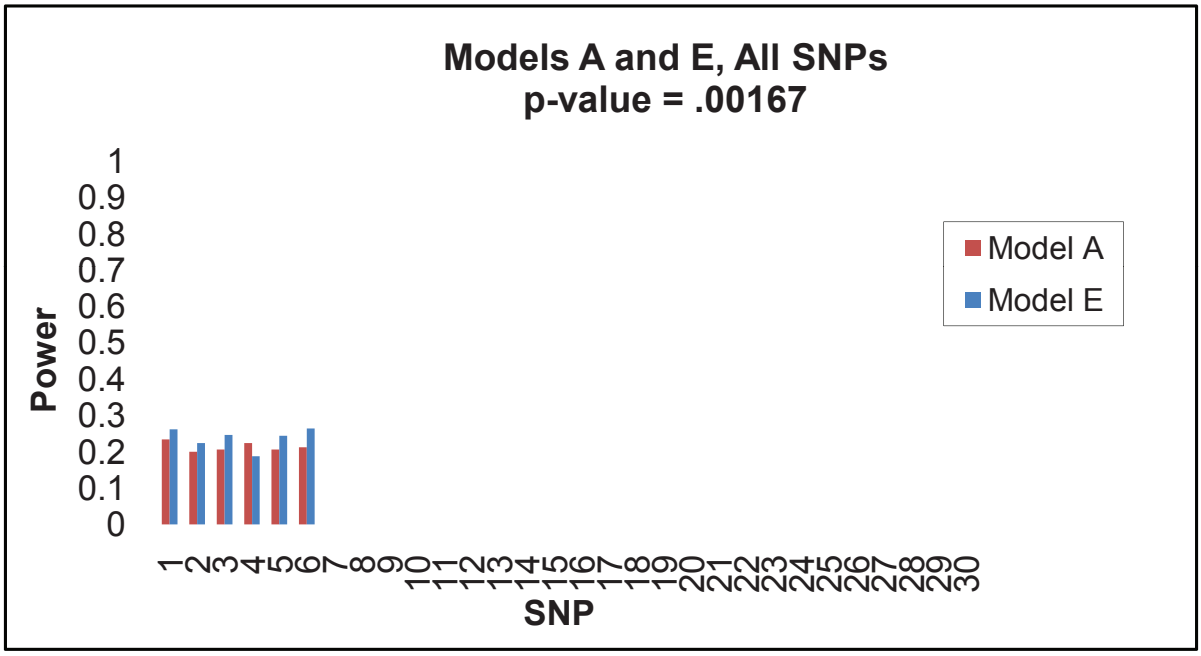

Figure 3: **MDR results for Additive (Model A) and Exact (Model E) simulations.** The results are for the Additive Model and the Exact Model. Displayed results are based on 500 observations per sample. The power is measured on the y-axis. The achieved power is shown for each of the 30 SNPs. The p-value is adjusted for the 30 SNPs. The power for RF for the two models is approximately 1 as seen in Supp. Data Fig. 1

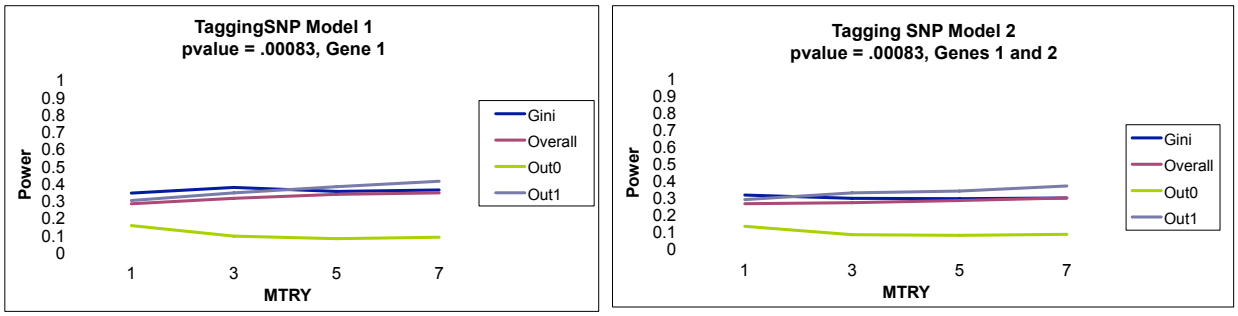

Figure 4: **Random Forests results for the Tagging SNPs Model 1 and 2 simulations.** The results are based on 500 observations. Power is measured on the y-axis and the four different values for mtry on the x-axis. The p-value is adjusted for the 60 SNPs, i.e. $0.05/60 = 0.00083$, and nT = 1000. The colored lines correspond to the achieved power for each of the four variable importance measures. For Model 1 only the results for Gene 1 are displayed. For Model 2 the average power over Genes 1 and 2 is displayed.

2 Simulations

2.1 Models A and E

Figures 1, 2, and 3.

2.2 Tagging SNPs

Figures 4 and 5, and Tables 1, 2, and 3.

3 Data Analysis

Data imputation. Missing values are not allowed in the R version of RF. Therefore, before beginning the data analysis, the missing SNP values were imputed with the expected allele count for each SNP based on the controls. As such, we defined the allele frequency for each SNP as: $f = (2 * N20 + N10)/(2 * N0)$, where $N20$ is the number of controls with genotype equal to 2, $N10$ is the number of controls with genotype equal to 1, and $N0$ is the total number of controls.

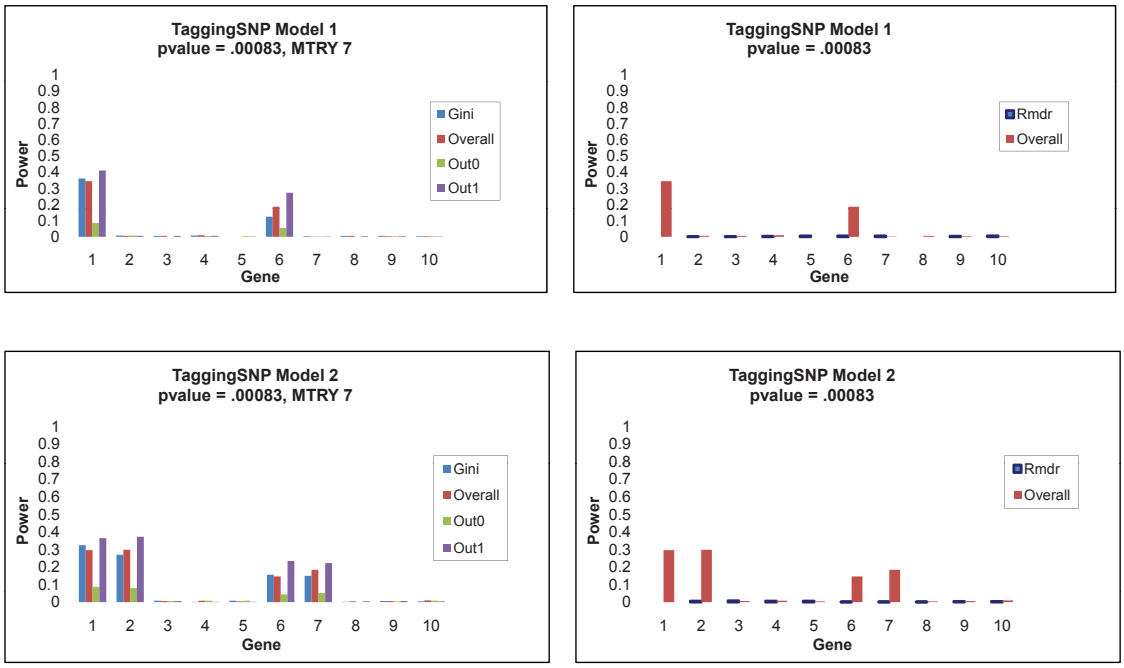

Figure 5: **Random Forest and MDR results for Tagging SNPs Model 1 and 2 simulations.** The results in the top row are based on Model 1 and the bottom row on Model 2. For each, the power is measured on the y-axis. The first column displays only results from RF. The second column compares RF to MDR. In the first, the power for RF (using only mtry = 7) for each of the four importance measures are shown across all 10 genes. In the second, the two algorithms are compared, with the overall importance measure and mtry = 7 for RF. For all shown results, there are 500 observation samples, (nT) is equal to 1000, and the p-value is adjusted for the 60 SNPs, i.e. $05/60 = 0.00083$.

Table 1: Haplotype Frequencies for GPX3

Causal SNP	Tagging SNPs	Freq
0	0 0 0 0 0 0	0.32107134298394
1	0 0 1 1 0 1	0.12036581986912-δ
0	0 1 0 0 0 0	0.09086890276393
0	0 0 0 0 0 1	0.07849771193634-δ/2
0	1 1 1 0 0 1	0.07216619820667
0	1 1 0 0 0 1	0.07083801403457
0	0 0 0 0 1 0	0.06100456469089-δ/2
0	0 1 1 0 0 1	0.05234293247470
0	1 1 0 0 0 0	0.04681030462982
0	1 0 0 0 0 0	0.03531025706472
0	0 0 1 0 0 0	0.02793906883632
0	0 1 0 0 0 1	0.02278488250898
0	0 0 1 1 0 1	δ
1	0 0 0 0 0 1	δ/2
1	0 0 0 0 1 0	δ/2

δ=0.0050, 0.0150, and 0.0260 for R^2=0.90, 0.75, and 0.60 respectively.

Table 2: Haplotype Frequencies for GPX4 (common gene scenario)

Causal SNP	Tagging SNPs	Freq
0	1 0 0 0 1 0	0.35061453
0	0 1 0 0 0 1	0.28190608
1	0 1 0 1 0 0	0.12741856-δ
0	1 0 0 0 0 0	0.06776808-δ/2
0	0 0 0 0 0 0	0.04074233
0	1 0 1 1 0 0	0.04006638-δ/2
0	0 0 0 0 1 0	0.03066169
0	0 1 0 0 1 0	0.02373138
0	0 1 0 0 0 0	0.02264183
0	1 0 0 0 0 1	0.01444912
1	1 0 0 0 0 0	δ/2
1	1 0 1 1 0 0	δ/2
0	0 1 0 1 0 0	δ

δ=0.0060, 0.0160, and 0.0300 for R^2=0.90, 0.75, and 0.60 respectively.

Table 3: Haplotype Frequencies for GPX4(rare gene scenario)

Causal SNP	Tagging SNPs	Freq
0	1 0 0 0 1 0	0.35061453
0	0 1 0 0 0 1	0.28190608
0	0 1 0 1 0 0	0.12741856
0	1 0 0 0 0 0	0.06776808
0	0 0 0 0 0 0	0.04074233
1	1 0 1 1 0 0	0.04006638-δ
0	0 0 0 0 1 0	0.03066169-δ/2
0	0 1 0 0 1 0	0.02373138-δ/2
0	0 1 0 0 0 0	0.02264183
0	1 0 0 0 0 1	0.01444912
0	1 0 1 1 0 0	δ
1	0 0 0 0 1 0	δ/2
1	0 1 0 0 1 0	δ/2

δ=0.0020, 0.0055, and 0.0093 for R^2=0.90, 0.75, and 0.60 respectively.