# Supplementary Material: A Novel Bayesian Graphical Model for Genome-Wide Multi-SNP Association Mapping

Yu Zhang

Department of Statistics, The Pennsylvania State University
421A Thomas Building, University Park, PA 16803
yuzhang@stat.psu.edu

# 1 Proof of Formula (3) in the Text

Given the disease status $Y$, a collection of selected SNPs $X_1$, and an undirected acyclic graph $G$ (consisting of cliques $C$ of SNPs and interaction $\Delta$ between cliques), we first construct a "directed" graph $G^*$ that has the same cliques and edges as in $G$, but the edges are directed. We do the following: 1) starting from an empty graph $G^* = \phi$, we randomly assign one clique $c_i \in C$ as a clique in $G^*$; 2) for every clique $c_i$ that is already in $G^*$, we sequentially assign new cliques, which have not been assigned to $G^*$ but are connected to $c_i$ (i.e., $\delta_{ij} = 1$), into $G^*$ in arbitrary orders, and we assign an edge from $c_i$ to $c_j$ in $G^*$; 3) If no more cliques in $G^*$ interact with the remaining unassigned cliques, we randomly assign one unassigned clique into $G^*$; and (4) we repeat steps (1-3) until all cliques are assigned into $G^*$. As a result, we obtain a "directed" graph $G^*$, which shares the same cliques and edges with $G$. Correspondingly, let $C^* = \{c_{[1]}, \cdots, c_{[K]}\}$ denote the $K$ cliques in $C$ that are partially ordered by the order they enter in $G^*$, and let $\Delta^* = \{\delta_{[i][j]}\}$ denote the interaction (directed edge) between $c_{[i]}$ and $c_{[j]}$.

Given $Y$ and $G^*$, we can use chain rules to define the joint distribution of SNPs in $X_1$ as a product of marginal and conditional probabilities:

$$\begin{aligned}\Pr(X_1|Y, I, G^*) &= \prod_{i=1}^{K}\left[\Pr(x_{c_{[i]}}|Y) \prod_{\{j:\delta_{[i][j]}=1, j>i\}} \Pr(x_{c_{[j]}}|x_{c_{[i]}}, Y)\right] \\ &= \prod_{i=1}^{K}\left[\Pr(x_{c_{[i]}}|Y) \prod_{\{j:\delta_{[i][j]}=1, j>i\}} \frac{\Pr(x_{c_{[i]}+c_{[j]}}|Y)}{\Pr(x_{c_{[i]}}|Y)}\right] \qquad (1)\end{aligned}$$

This is a factorization of Bayesian network models, and is a valid probability function because $G$ (and $G^*$) is acyclic.

It is easily checked that $\Pr(x_{c_{[i]}}|Y)$ for each clique $c_{[i]}$ appears exactly once in the numerator of (1), and exactly $k_i$ times in the denominator, where $k_i$ denotes the number of cliques connected with $c_{[i]}$. Furthermore, $\Pr(x_{c_{[i]}+c_{[j]}}|Y)$ for each pair of interacting cliques appears exactly once. Since we construct $G^*$ arbitrarily, it follows that formula (1) is invariant to the order of cliques and the direction of edges presented in $G^*$. Therefore, we can rewrite formula (1) in a simpler form as

$$\Pr(X_1|Y, G) = \prod_{i=1}^{K} \Pr(x_{c_i}|Y) \prod_{\{i,j:\delta_{ij}=1\}} \frac{\Pr(x_{c_i}, x_{c_j}|Y)}{\Pr(x_{c_i}|Y)\Pr(x_{c_j}|Y)} \qquad (2)$$

which yields the formula (3) in the text.

# 2 Additional Results

## 2.1 Rank Power using 50kb Windows

Figure S1 shows the rank power comparison of seven methods on the 10,000 SNP datasets. The calculation is similar to that shown in Figure1 in the main text, but a disease SNP is treated as detected if there is a top ranked SNP within 50kb to the disease SNP. Figure S2 shows the rank power of different methods on the 100,000 SNP datasets. Again, a disease SNP is detected if there is a top ranked SNP within 50kb to the disease SNP.

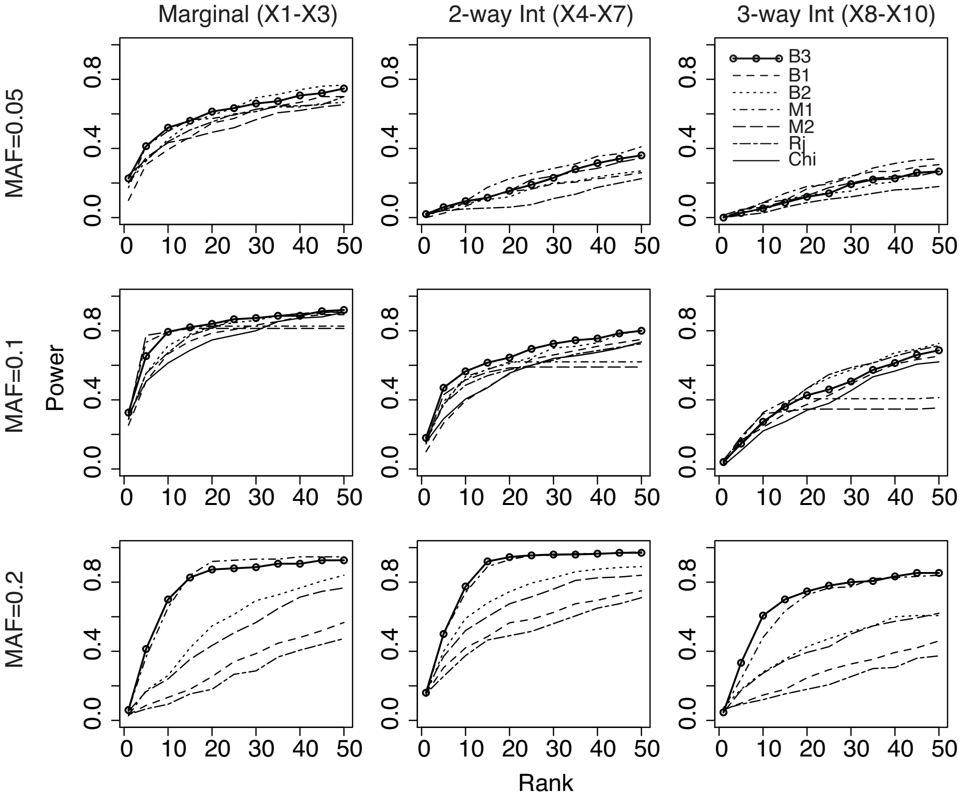

Figure S1: Rank power of BEAM3(B3, solid line with circle), BEAM1 (B1, dashed), BEAM2 (B2, dotted), Mendel-Single (M1, dotdash), Mendel-Pairwise (M2, longdash), RandomJungle (Rj, twodash), and ChiSq (Chi, solid black) on the 10,000 SNP datasets. A disease SNP is captured if within its 50kb neighborhood there is at least one top ranked SNPs.

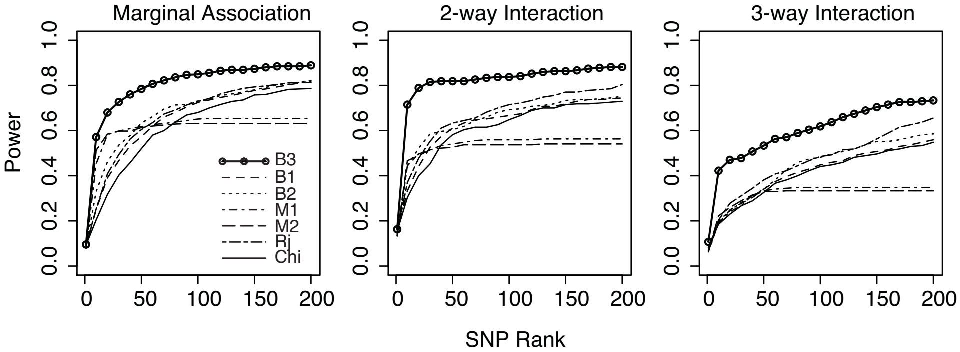

Figure S2: Rank power of BEAM3(B3, solid line with circle), BEAM1 (B1, dashed), BEAM2 (B2, dotted), Mendel-Single (M1, dotdash), Mendel-Pairwise (M2, longdash), RandomJungle (Rj, twodash), and ChiSq (Chi, solid black) on the 100,000 SNP datasets. A disease SNP is captured if within its 50kb neighborhood there is at least one top ranked SNPs.

## 2.2 Estimated Number of Disease SNPs

Table S1 shows the estimated number of disease SNPs by each method in simulated datasets with 10,000 SNPs and disease MAF=0.05. Table S2 shows the same calculation in simulated datasets with 10,000 SNPs and disease MAF=0.1. In these datasets, the association signals are weak, and thus the estimation is confounded by two factors: 1) missing a true disease SNP; and 2) over estimate the number of disease SNPs. For example, it appeared that BEAM2 estimated the most accurate number of marginal SNPs in Table S1, first row, and Mendel-Single estimated the most accurate number of marginal SNPs in Table S2, first row. By checking their selected SNPs, however, we found that often multiple selected SNPs corresponded to just one disease SNP, and thus the selected SNPs are redundant.

Table S1: Estimated numbers of disease SNPs by BEAM3 (B3), BEAM1 (B1), and BEAM2 (B2), and the numbers of SNPs selected by Mendel-Single (M1), Mendel-Pairiwise (M2), RandomJungle (Rj), and ChiSq (Chi). Disease MAF=0.05.

| dSNPs | True Size | B3 | B1 | B2 | M1 | M2 | Rj | Chi |
|---|---|---|---|---|---|---|---|---|
| $X_1 \sim X_3$ | Single: 3 | 0.74 | 1.79 | **3.03**[a] | 0.90 | 0.06 | 0.54 | 4.10 |
| (Region | | (0.64)[b] | (3.29) | (3.43) | (1.07) | (0.31) | (1.47) | (11.85) |
| 1, 2) | Interact: 0 | **0.05** | 0.22 | 0.76 | n/a | 0.62 | n/a | n/a |
| | | (0.26) | (0.61) | (0.98) | n/a | (1.19) | n/a | n/a |
| $X_4 \sim X_{10}$ | Single: 0 | 0.21 | 0.67 | 1.24 | 0.06 | **0.00** | 0.02 | 0.60 |
| (Region | | (0.19) | (2.09) | (2.62) | (0.24) | (0.00) | (0.14) | (4.10) |
| 3,4,5) | Interact: 7 | 0.16 | **0.82** | 0.34 | n/a | 0.06 | n/a | n/a |
| | | (0.52) | (1.01) | (0.80) | n/a | (0.31) | n/a | n/a |

[a]: most accurate estimation; [b]: standard deviation.

Table S2: Estimated numbers of disease SNPs by BEAM3 (B3), BEAM1 (B1), and BEAM2 (B2), and the numbers of SNPs selected by Mendel-Single (M1), Mendel-Pairiwise (M2), RandomJungle (Rj), and ChiSq (Chi). Disease MAF=0.1.

| dSNPs | True Size | B3 | B1 | B2 | M1 | M2 | Rj | Chi |
|---|---|---|---|---|---|---|---|---|
| $X_1 \sim X_3$ | Single: 3 | 1.67 | 5.89 | 9.68 | **3.14**[a] | 0.28 | 2.02 | 23.1 |
| (Region | | (0.66)[b] | (8.51) | (8.59) | (1.95) | (0.70) | (2.56) | (33.6) |
| 1, 2) | Interact: 0 | **0.04** | 0.28 | 0.74 | n/a | 3.02 | n/a | n/a |
| | | (0.08) | (0.70) | (1.02) | n/a | (1.73) | n/a | n/a |
| $X_4 \sim X_{10}$ | Single: 0 | 0.47 | 2.59 | 5.09 | 1.54 | **0.12** | 0.72 | 8.16 |
| (Region | | (0.34) | (4.79) | (7.49) | (2.22) | (0.44) | (1.86) | (17.76) |
| 3,4,5) | Interact: 7 | 1.60 | 0.84 | 1.31 | n/a | **2.02** | n/a | n/a |
| | | (1.81) | (0.97) | (1.20) | n/a | (1.95) | n/a | n/a |

[a]: most accurate estimation; [b]: standard deviation.

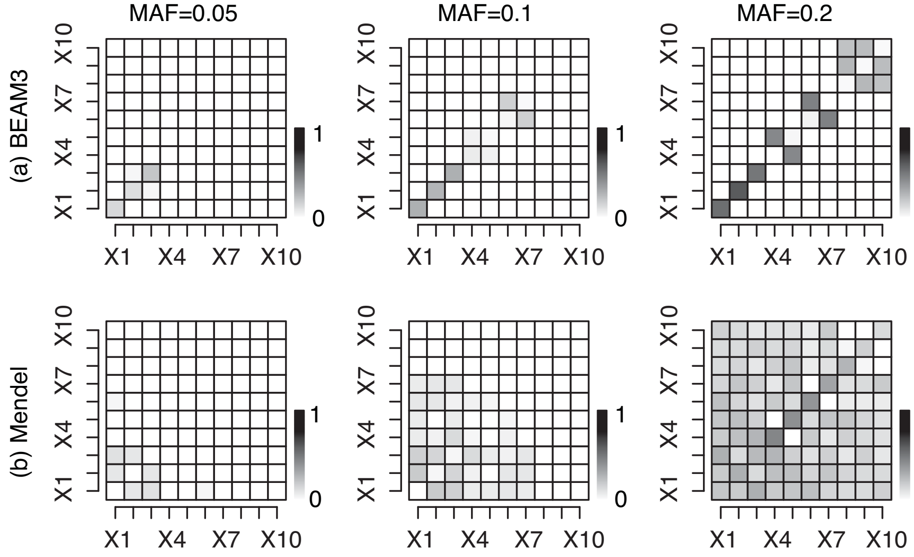

Figure S3: Inferred interaction structures by (a) BEAM3 and (b) Mendel-Pairwise on 50 datasets of 10,000 SNPs, with disease MAF=0.05, 0.1, 0.2, respectively. The results are calculated based on 5kb window.

### 2.3 Inferred Disease Structure: Main Effect versus Interaction

Heatmap of the identified disease structures by BEAM3 and Mendel. Figure S3 is similar to Figure 3 in the main text, but the results are obtained by using a 5kb window. That is, a disease SNP (or a pairwise interaction) is detected if within 5kb to the true disease SNPs, there is a association (or interaction) detected by each method. Figure S4 is similar to Figure 4b in the main text, for the 100,000 SNPs, but again using a 5kb window.

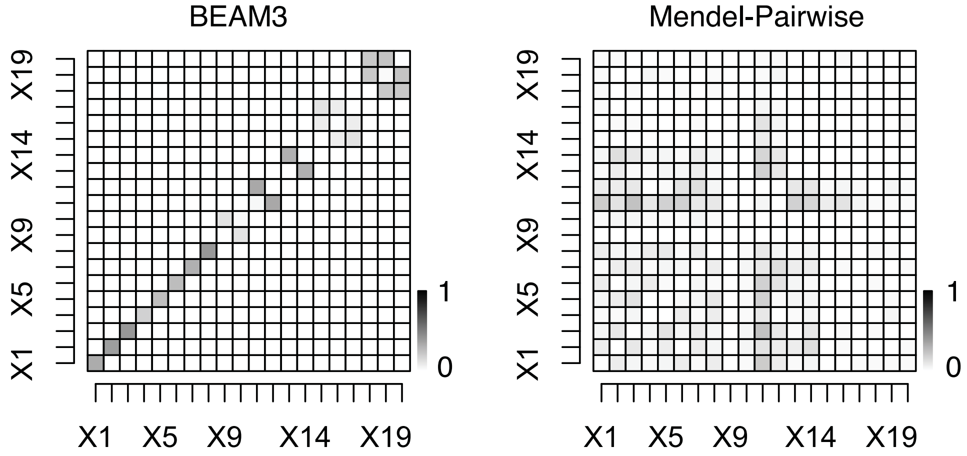

Figure S4: Inferred interaction structures by BEAM3 and Mendel-Pairwise from 50 datasets of 100,000 SNPs. The results are calculated based on 5kb window.