University of California

Los Angeles

Bayesian Inference on Allele Group Structure for High Order Interactions in Genome-Wide Association Studies

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Albert Wong

2013

Abstract of the Thesis

Bayesian Inference on Allele Group Structure for High Order Interactions in Genome-Wide Association Studies

by

Albert Wong

Master of Science in Statistics

University of California, Los Angeles, 2013

Professor Qing Zhou, Chair

Sophisticated Bayesian methods are often used to identify a collection of alleles that are jointly associated with a particular disease. A disease might not be expressed when only one of these alleles is present, but each associated allele might interact with each other in a rather complicated way, causing a disease to be expressed. In investigating a patient's susceptibility to a disease, it is often useful to group the collection of associated alleles according to their risk factors.

Our goal is to find the most likely grouping structure of alleles $C_1, ..., C_m$ associated with Rheumatoid Arthritis given a case-control data. The number of ways to group these m alleles is given by the m^{th} Bell number B_m, defined recursively by $B_m = \sum_{k=0}^{m-1} \binom{m-1}{k} B_k$ with $B_0 = B_1 = 1$. For 10 alleles, this translates to 115,975 groupings. For $m = 15$, we have over a billion ways to group $C_1, ..., C_m$. Clearly computing the probability for each grouping soon become intractable. A combination of Metropolis-Hastings and local search algorithm is proposed to accomplish this task. This strategy is first implemented on simulated data, with a sufficiently large sample size and a known grouping structure, and the correct grouping is obtained. Stable results are obtained as the algorithm is run multiple times on Rheumatoid Arthritis data.

The thesis of Albert Wong is approved.

Nicolas Christou

Ying Nian Wu

Qing Zhou, Committee Chair

University of California, Los Angeles

2013

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Relevant Information

Rheumatoid Arthritis (RA) is a common chronic inflammatory autoimmune disease that leads to a progressive joint destruction. With the advances of genotyping technology, genetic information associated with RA becomes available from the Genome-Wide Association Studies (GWAS) data. GWAS is an examination of common genetic variants in different individuals to determine if any variant is associated with a trait or disease. In contrast to other studies that usually consider only one or few genetic locations, GWAS examine each individual's entire genetic information.

GWAS typically compare the DNA of two groups: cases and controls. DNA is the basic information molecule that encodes genetic instructions used in the development of a living organism. DNA is composed of 4 types of bases: adenine (A), guanine (G), cytosine (C) and thymine (T). A Single-Nucleotide Polymorhism (SNP) is a DNA sequence variation caused by a variation in a single nucleotide found in the members of population. For example, ATC and ATG might have been found in different individuals and we say that 2 alleles are observed. In the case when one variant is more common in the people with a trait, we then associate such variant with the trait.

1.2 Structure of the Thesis

The rest of the thesis consists of 6 chapters. Chapter 2 gives an explanation of the data and some relevant works. Statistical methods used in this thesis are described in chapter 3. In chapter 4, described statistical methods are applied first to simulated data and then to RA data. Conclusion of the thesis is found in chapter 5. Chapter 6 provides a detailed derivation of the posterior computation described in chapter 3 and lists a few more assumptions that are needed to make the posterior computation possible.

CHAPTER 2

Data

GWAS data used in this thesis are obtained from the North American Rheumatoid Arthritis Consortium (NARAC). Multiple locations in DNA sequence associated with RA susceptibility have been indentified by genome-wide association studies (Cornelis, F. et al., 1998, WTCCC, 2007, Plenge, R.M. et al., 2007, Stahl, E.A. et al., 2010). The presence of a single genetic variant (allele) found in these locations might not cause RA to be expressed but they might interact with each other in a complex way, causing RA to be expressed (Manolio, T.A. et al., 2009, Wu, Z., Zhao, H., 2009). Many interactions (collections of alleles) associated with RA are detected by Jing Zhang et al. (2012) who performed the first genome-wide high order interaction analysis for RA using Bayesian epistasis association mapping (BEAM and BEAM2) methods (Zhang, Y., Liu, J.S., 2007, Zhang, Y. et al., 2011). From the 90 SNPs involved in the 319 interactions identified, 18 SNPs that have good genotype data quality are retreived from the GWAS data from NARAC. The full version tables of high order interactions representatives and associated SNP annotations could be found in Jing Zhang et al. (2012).

In this thesis we look at 7 significant high order interactions. A set of 3 SNPs, each SNP assumes a value of 0, 1 or 2, is associated with each interaction studied. Every possible configuration of these three SNPs is considered as an allele, giving each interaction a set of 3^3 alleles. The goal is to make inference on the most likely allele grouping structure for each of the 7 interactions. Given $C_1, ..., C_{27}$ alleles for each interaction, data is read from a 2 by 27 matrix. The first row summarizes the number of patients in the diseased group that have C_i. The second row of the matrix records the counts from the control group. There are a total of 2,002 subjects (862 cases from the diseased pool and 1,140 from control).

interaction	Pool	000	001	002	010	011	012	020	021	022	100	101	102	110	111
26	diseased	94	109	21	54	23	1	10	2	0	199	163	24	87	77
	control	116	38	3	74	16	1	4	1	0	201	79	0	103	24
31	diseased	28	7	1	114	20	1	97	14	5	125	39	3	254	64
	control	90	11	1	204	43	4	141	29	2	103	21	1	130	22
52	diseased	295	69	4	278	54	2	55	6	1	151	36	2	103	33
	control	347	68	8	122	22	1	4	1	0	181	31	2	42	8
180	diseased	62	9	1	160	36	1	173	38	2	140	30	4	253	61
	control	47	8	1	78	18	2	41	12	1	173	38	3	211	38
255	diseased	763	159	10	33	12	0	1	0	0	48	8	1	62	18
	control	497	97	6	107	17	4	3	0	0	15	2	0	72	12
303	diseased	99	186	77	57	153	116	8	19	32	60	97	25	39	73
	control	289	121	6	39	97	3	1	2	2	156	57	0	22	26
307	diseased	295	69	4	278	54	2	55	6	1	151	36	2	103	33
	control	347	68	8	122	22	1	4	1	0	181	31	2	42	8

Table 2.1: Allele counts for each of the 7 significant interactions. Each allele has two counts, one from each diseased and control pools.

interaction	Pool	112	120	121	122	200	201	202	210	211	212	220	221	222
26	diseased	3	13	6	0	75	62	17	48	38	2	8	2	0
	control	1	11	3	0	106	28	2	37	10	0	1	2	0
31	diseased	2	53	6	0	159	40	1	61	11	0	8	0	0
	control	4	24	5	0	7	0	0	2	0	0	1	1	0
52	diseased	2	5	1	0	27	3	1	7	2	1	0	0	0
	control	0	0	2	0	13	2	1	4	2	0	0	0	0
180	diseased	2	19	5	1	90	19	2	14	3	0	2	0	0
	control	4	8	1	0	143	18	1	10	1	0	0	0	0
255	diseased	2	0	1	0	0	0	0	2	1	0	3	0	0
	control	2	10	3	0	0	0	0	1	0	0	2	1	0
303	diseased	24	3	3	0	13	12	6	6	3	0	0	0	0
	control	0	0	2	0	12	4	0	5	1	0	0	0	0
307	diseased	2	5	1	0	27	3	1	7	2	1	0	0	0
	control	0	0	2	0	13	2	1	4	2	0	0	0	0

Table 2.2: Allele counts for each of the 7 significant interactions. Each allele has two counts, one from each diseased and control pools (continued).

CHAPTER 3

Methods

3.1 Metropolis-Hastings Algorithm

Metropolis-Hastings Algorithm is a Markov Chain Monte Carlo method useful in generating samples from a complicated probability distribution $P(\cdot)$. This is achieved by constructing a reversible Markov Chain that has $P(\cdot)$ as its stationary distribution. In order to govern the movement of the Markov Chain, a transition probability that is allowed to depend on the current state x is specified and denoted by $q(x, \cdot)$. We refer this as the proposal probability for the Metropolis-Hastings Algorithm. The probability of making the proposed move to y from x is then computed using $\alpha(x, y) = \min\left\{\frac{P(y)}{P(x)}\frac{q(y,x)}{q(x,y)}, 1\right\}$. We refer this as acceptance probability of the Metropolis-Hastings Algorithm. It is clear that the probability of making a move from x to y is $P_{MH}(x, y) = q(x, y)\alpha(x, y)$. One could check that detailed balance condition $P(x)P_{MH}(x, y) = P(y)P_{MH}(y, x)$ holds and the Markov Chain constructed converges to the distribution $P(\cdot)$.

Even though the constructed Markov Chain converges to the prespecified distribution, we still need to know from which point on it is safe to consider the constructed Markov Chain as a genuine sample. Burn-in period is the time it takes for the constructed Markov Chain to approximately converge. In practice, burn-in period is often determined by running the algorithm a couple of times. Every state visited after the burn-in period is a genuine, though correlated, sample from distribution $P(\cdot)$.

Here is the procedure to run Metropolis-Hastings Algorithm with an initial value $x_{(0)}$. Suppose that we have generated $x_{(t)}$, then:

- Generate y from $q(x_{(t)}, \cdot)$
- Compute $\alpha(x_{(t)}, y) = \min\left\{\frac{P(y)}{P(x_{(t)})}\frac{q(y,x_{(t)})}{q(x_{(t)},y)}, 1\right\}$
- Generate u from Uniform(0,1)
- If $u \leq \alpha(x_{(t)}, y)$, set $x_{(t+1)} = y$
- Else, set $x_{(t+1)} = x_{(t)}$

3.2 Maximum A Posteriori (MAP) Estimate

We could find the MAP estimate of an unknown parameter when we are working in Bayesian framework. First we define two common terminologies used in Bayesian statistics: prior and posterior distributions of parameter being estimated. Let x be our data and θ be the parameter to be estimated after observing x.

Prior distribution of θ is simply a probability distribution $P(\theta)$ assigned to θ. This assignment of probability distribution $P(\theta)$ means that θ is treated as a random variable, as opposed to a fixed yet unknown constant in frequentists approach. Prior $P(\theta)$ reflects our current knowledge on θ. In the case when we don't know any information about θ, we set $P(\theta) \propto 1$, called uniform prior, reflecting our ignorance or lack of information. Posterior distribution of θ is $P(\theta|x)$, i.e. conditional distribution of θ given that data x is observed.

As the name suggests, MAP estimate of the unknown parameter θ is simply $\theta*$ such that

$$\theta* = \arg\max_{\theta \in \Theta} P(\theta|x)$$

3.3 MAP estimate of allele group structure using Metropolis-Hastings Algorithm and local search

Consider m alleles $C_1, ..., C_m$ that are associated with a disease. For each allele we have counts from a pool of diseased patients and a pool of control patients. So we have $n_1, ..., n_m$ and $n'_1, ..., n'_m$, where they represent the counts from diseased and control groups, respectively. It is proposed that $C_1, ..., C_m$ could be partitioned into L groups, $1 \leq L \leq m$. Such a grouping is labeled $(S_1, ..., S_L)$, where $S_i = \{C_{i1}, ..., C_{i|S_i|}\}$ and every allele in S_i shares the same risk factor. Our goal is to make inferences on both L and the grouping $(S_1, ..., S_L)$ given the data $n_1, ..., n_m$ and $n'_1, ..., n'_m$. In order to achieve this, we draw a bunch of samples $(S_1, ..., S_L)$ from the posterior distribution $P(\cdot|n_1, ..., n_m, n'_1, ..., n'_m)$ using Metropolis-Hastings Algorithm. We then find the grouping that has the maximum posterior probability among the ones being sampled and do a local search starting from that point.

In order to run Metropolis-Hastings Algorithm, we need to perform two computations at each iteration. Firstly, we should be able to compute the posterior probability $P(S_1, ..., S_L|n_1, ..., n_m, n'_1, ..., n'_m)$ for each grouping $(S_1, ..., S_L)$ up to a proportionality constant. Secondly, we also need to specify and compute the proposal probability that governs the movement from one state to the next. Proposal probability is denoted by $q(x, y)$ where x and y are two particular groupings of $C_1, ..., C_m$.

Bayesian framework is specified to achieve the first computation. Suppose that a particular grouping $(S_1, ..., S_L)$ is fixed. Let $n_{S_i} = \sum_j^{|S_i|} n_{ij}$, where $|S_i|$ is the number of alleles in group i, and n_{ij} corresponds to the count of allele C_{ij} found in the patients with disease. $n'_{S_i} = \sum_j^{|S_i|} n'_{ij}$ is similarly defined for the control group. We assume the following model is true:

$$(n_{i1}, ..., n_{i|S_i|}) \sim \text{Multinomial}(n_{S_i}, \vec{\theta}_i) \tag{3.1}$$

$$(n'_{i1}, ..., n'_{i|S_i|}) \sim \text{Multinomial}(n'_{S_i}, \vec{\theta}_i) \tag{3.2}$$

$$(n_{S_1}, ..., n_{S_L}) \sim \text{Multinomial}(\sum_i^m n_i, \vec{p}) \tag{3.3}$$

$$(n'_{S_1}, ..., n'_{S_L}) \sim \text{Multinomial}(\sum_i^m n'_i, \vec{p}') \tag{3.4}$$

where $\vec{\theta}_i = (\theta_{i1}, ..., \theta_{i|S_i|})$, $\vec{p} = (p_1, ..., p_L)$, and $\vec{p}' = (p'_1, ..., p'_L)$. The first two equations suggest that the probability distributions for allele counts within any group S_i are identical in both diseased and control cases. However, we allow the probability distributions p and p' for $(n_{S_1}, ..., n_{S_L})$ and $(n'_{S_1}, ..., n'_{S_L})$, respectively, to be different. This assumption makes sense since the presence of certain group, say S_i, might be associated with the presence or absence of the disease in question. We also need to specify the priors for the parameters $\vec{\theta}_i, \vec{p}$ and $\vec{p}'$:

$$\vec{\theta}_i|S_i \sim \text{Dirichlet}(\frac{1}{|S_i|}, ..., \frac{1}{|S_i|}) \tag{3.5}$$

$$\vec{p}|L, \vec{p}'|L \sim \text{Dirichlet}(\alpha_1, ..., \alpha_L) \tag{3.6}$$

where $\alpha_i = \frac{1}{L}, i = 1, ..., L$.

At this point, we are ready to compute posterior probability for each grouping. Applying Bayes' rule, we get:

$$P(S_1, ..., S_L|n_1, ..., n_m, n'_1, ..., n'_m) \propto P(S_1, ..., S_L)P(n_1, ..., n_m, n'_1, ...n'_m|S_1, ..., S_L) \tag{3.7}$$

We further assume that $P(S_1, ..., S_L) \propto \gamma^{-L}, \gamma \geq 1$. This means that we penalize groupings with large L. For convenience however, γ is set to 1 in the algorithm. By (3.7), it is enough to compute the likelihood $P(n_1, ..., n_m, n'_1, ..., n'_m|S_1, ..., S_L)$.

After integrating $\vec{\theta}_i, \vec{p}$ and $\vec{p}'$ out, we get:

$$\begin{aligned} &P(n_1, ..., n_m, n'_1, ..., n'_m | S_1, ..., S_L) \\ &= \frac{1}{\Gamma(1+n)} \prod_{i=1}^{L} \frac{\Gamma(n_{S_i} + \frac{1}{L})}{\Gamma(\frac{1}{L})} \times \frac{1}{\Gamma(1+n')} \prod_{i=1}^{L} \frac{\Gamma(n'_{S_i} + \frac{1}{L})}{\Gamma(\frac{1}{L})} \\ &\prod_{i=1}^{L} \frac{1}{\Gamma(1+n_{S_i}+n'_{S_i})} \times \prod_{i=1}^{L} \prod_{j=1}^{|S_i|} \frac{\Gamma(n_{ij} + n'_{ij} + \frac{1}{|S_i|})}{\Gamma(\frac{1}{|S_i|})} \end{aligned} \tag{3.8}$$

where $n = \sum_{i=1}^{L} n_{S_i} = \sum_{j=1}^{m} n_j$ and $n' = \sum_{i=1}^{L} n'_{S_i} = \sum_{j=1}^{m} n'_j$. The full derivation of the likelihood is shown in the appendix.

Having the formula for computing posterior probability for each grouping, we still need to specify the proposal probability of the Metropolis-Hastings Algorithm at each iteration. To simplify our notations, we would write states $x = (S_1, ..., S_L)$ and $y = (S'_1, ..., S'_{L'})$ and assume that y is accessible from x everytime we write a pair (x, y). Let $q(x, y)$ denote the probability of proposing a move to y from x. Before we could define $q(x, y)$, we first need to introduce three different update techniques:

- Combine 2 randomly chosen groups: We randomly choose i and j and collapse S_i and S_j into one group. Probability of making such a move is $q_1(x, y) = \frac{1}{\binom{L}{2}}$, where $L =$ the number of groups in state x.

- Split a randomly chosen group into two: We first randomly choose one among the splittable groups in state x. Suppose j^{th} group is selected, and we then split it randomly. In this case we have $q_2(x, y) = \frac{1}{|\{S_i : |S_i| \geq 2\}|} \frac{1}{\frac{2^{|S_j|}-2}{2}}$. Notice that $\frac{2^{|S_j|}-2}{2}$ is the total number of ways one could split a group containing $|S_j|$ alleles into two.

- Switch groups of two alleles: First we randomly choose two groups, say S_i and S_j, in x. We then randomly pick 2 alleles, one from each groups, and swap the two. In this case we notice that $q_3(x, y) > 0$ if and only if $q_3(y, x) > 0$. Moreover, we also have $\frac{q_3(y,x)}{q_3(x,y)} = 1$.

The proposal probability of the Metropolis-Hastings Algorithm $q(x, y)$ is then defined by:

$$q(x, y) = \frac{1}{k} q_j(x, y),\ k, j = 1, 2, 3 \tag{3.9}$$

where $k = k_x = |\{l = 1, 2, 3 : \text{there exists } y_l \text{ such that } q_l(x, y_l) > 0\}|$, i.e. it denotes how many different update techniques are available if we were at state x. $j = j_{x,y} \in \{l = 1, 2, 3 : q_l(x, y) > 0\}$. Given the pair (x, y), j is unique since no two different update techniques would lead x to the same destination y.

We finish the inference on the grouping of $C_1, ..., C_m$ by doing a local search to make sure that the grouping that corresponds to at least a local maximum of the posterior distribution $P(S_1, ..., S_L|n_1, ..., n_m, n'_1, ..., n'_m)$ is found. Suppose that genuine samples from posterior distribution are already obtained by running Metropolis-Hastings Algorithm, we then can find the grouping with maximal posterior among the ones generated. Starting from this point we can start the local search by using an algorithm similar to the Metropolis-Hastings. Let $q(x, y)$ denote the same proposal probability described above. Acceptance probability is modified so that a new state is accepted if and only if there is an increase in the posterior distribution. Once this is done, we would obtain a MAP estimate of the unknown parameter $(S_1, ..., S_L)$.

3.3.1 Convergence in distribution of the constructed Markov Chain to the posterior

In this section, we show the convergence of the constructed Markov Chain to the posterior distribution by showing that the transition probability of the Metropolis-Hastings algorithm $P_{MH}(x, y) = q(x, y)\alpha(x, y)$ satisfies detailed balance condition. As before, we denote any two particular groupings of $C_1, ..., C_m$ with x and y. Furthermore, conditional upon data is assumed throughout and we write posterior $P(x|n_1, ..., n_m, n'_1, ..., n'_m) = \pi(x)$. Detailed balance condition for the transition probability of the Metropolis-Hastings Algorithm is then described by:

$$\pi(x)q(x,y)\alpha(x,y) = \pi(y)q(y,x)\alpha(y,x)\forall x,y \tag{3.10}$$

Summing over y on both sides of (3.10) shows that detailed balance implies convergence in distribution to π.

Before showing that detailed balance condition is satisfied, we make two observations. Firstly, we have $\pi(x) > 0$ for any x. This is true by equations (3.7) and (3.8). The second observation is that $q(x,y) > 0$ if and only if $q(y,x) > 0$. In order to show that this is the case, we assume without loss of generality that $q(x,y) > 0$. This implies that there is unique j such that $q_j(x,y) > 0$. This, in turn, implies that there exists a j' such that $q_{j'}(y,x) > 0$. This is true since if y were obtained from x by collapsing two groups in x into one then x is obtained from y by splitting that new big group. Following the same logic, x must have been obtained from y by collapsing if y were obtained from x by splitting. We have $j = j' = 3$ if y were obtained from x by swapping. Hence we have $q(y,x) > 0$ whenever $q(x,y) > 0$.

Now we are ready to show that detailed balance condition is satisfied. If $q(x,y) = 0$ then there is nothing to show since both sides of equation (3.10) would be zero by the second observation above. So we assume that $q(x,y)$, and hence $q(y,x)$, is positive. We further assume that $\alpha(x,y) = 1$, i.e., $\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} > 1$. We then have the left hand side of (3.10) equal to $\pi(x)q(x,y)$, while the right hand side of (3.10) is equal to $\pi(y)q(y,x)\alpha(y,x) = \pi(y)q(y,x)\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)} = \pi(x)q(x,y)$. The case when $\alpha(x,y) < 1$ is treated similarly.

CHAPTER 4

Analysis

4.1 Analysis on Generated Data

In order to test the correctness of the implementation of the algorithm used in the thesis, we first run it on a simulated data. We would refer the algorithm that combines Metropolis-Hastings and local search as the MAP algorithm. First we fix a grouping structure and hope that we could find the correct grouping by running the MAP algorithm. Equations (3.1)-(3.4) show how data $(n_1, ..., n_m, n'_1, ..., n'_m)$ is generated given a fixed grouping $(S_1, ..., S_L)$.

Suppose that we have 10 alleles $\{C_1, ..., C_{10}\}$ and a fixed grouping $(S_1, ..., S_4) = (\{C_1, C_2, C_3\}, \{C_4, C_5\}, \{C_6, C_7, C_8\}, \{C_9, C_{10}\})$. After specifying the total number of patients for each diseased and control pools, data $(n_1, ..., n_{10}, n'_1, ..., n'_{10})$ is then generated according to the following frequencies.

Grouping	Disease Frequency	Control Frequency	Allele Distribution
$\{C_1, C_2, C_3\}$	$p_1 = 0.5$	$p'_1 = 0.2$	$\theta_{11} = 0.3, \theta_{12} = 0.3, \theta_{13} = 0.4$
$\{C_4, C_5\}$	$p_2 = 0.2$	$p'_2 = 0.1$	$\theta_{21} = 0.6, \theta_{22} = 0.4$
$\{C_6, C_7, C_8\}$	$p_3 = 0.1$	$p'_3 = 0.2$	$\theta_{31} = 0.5, \theta_{32} = 0.2, \theta_{33} = 0.3$
$\{C_9, C_{10}\}$	$p_4 = 0.2$	$p'_4 = 0.5$	$\theta_{41} = 0.5, \theta_{42} = 0.5$

Table 4.1: Parameters $\vec{p}, \vec{p}', \vec{\theta}_i$ used in data generation

If the MAP algorithm were to be implemented correctly, we should be able to find the exact grouping $(\{C_1, C_2, C_3\}, \{C_4, C_5\}, \{C_6, C_7, C_8\}, \{C_9, C_{10}\})$ when we have a sufficiently large sample size for both diseased and control pools. The algorithm is used only to find the MLE estimate of parameter $(S_1, ..., S_L)$ as data $(n_1, ..., n_{10}, n'_1, ..., n'_{10})$ is observed. This is true since (3.7) and $P(S_1, ..., S_L) \propto 1$ imply:

$$\arg\max_{(S_1,...,S_L)} P(S_1, ..., S_L | n_1, ..., n_{10}, n'_1, ..., n'_{10}) =$$
$$\arg\max_{(S_1,...,S_L)} P(n_1, ..., n_{10}, n'_1, ..., n'_{10} | S_1, ..., S_L)$$

We note that the right hand side of the above equation solves for maximum likelihood estimate of parameter $(S_1, ..., S_L)$.

Using the model described in (3.1)-(3.4) and parameters in Table 4.1, we generate three sets of data with total counts of 1000, 10000, and 100000 for each diseased and control cases. Here is the realization of the counts used in testing our MAP algorithm:

Counts	Pool	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
1,000	diseased	134	138	201	124	77	56	15	25	112	118
	control	46	67	78	61	32	97	41	64	283	231
10,000	diseased	1536	1486	1987	1282	797	508	206	272	1022	904
	control	624	589	823	598	404	992	408	614	2474	2474
100,000	diseased	1529	15055	19900	12009	7964	4960	2024	3007	9995	9957
	control	6111	6015	8192	5999	3992	10035	4043	5919	24869	24865

For each generated data set, we run the Metroplis-Hastings algorithm for 100,000 iterations and remove the first 1,000 samples to allow for the burn-in period. Starting from the grouping with maximum posterior among all the groupings sampled by Metropolis-Hastings algorithm, we then perform a local search with the same transtition probability. A new grouping is accepted if and only if there is an increase in posterior distribution. Table 4.2 shows that MAP algorithm finds the actual grouping $(S_1, S_2, S_3, S_4) = (\{C_1, C_2, C_3\}, \{C_4, C_5\}, \{C_6, C_7, C_8\}, \{C_9, C_{10}\})$ when we have a decent size of both diseased and control groups.

In addition to identifying the grouping with maximal posterior for each generated data set, we also record how many times alleles i and j appear together among all the groupings sampled by the Metropolis-Hastings algorithm in a 10 by 10 matrix. Diagonal entries are ignored as we should have all ones. Each entry is either very close to 0 or very close to 1 as we have large iterations and large total counts. Each entry in this 10 by 10 symmetric matrix (after putting all the ones in the diagonal) is then mapped into a set of 5 colors. Entries with high probabilities are mapped to colors with low intensity (white) while entries with low probabilities are mapped to colors with high intensiy (red). Reordering of alleles in the heat map is done automatically to facilitate visualization of the grouping structure.

Finally, we try to construct a grouping from this 10 by 10 matrix and compare it to the grouping obtained by running the MAP algorithm. A cutoff point of 0.5 is specified. This means that a line is drawn between alleles i and j whenever the entry in the i^{th} row and j^{th} column is greater than 0.5. Note that each grouping constructed this way would be different if we were to be stricter by increasing the cutoff point. For example, if we were to increase the cutoff point to 0.8, then the grouping with maximal posterior and its reconstruction coincide only when we have 100,000 samples.

Count	Maximum a posteriori estimate
1,000	$\{C_1, C_2, C_3, C_4, C_5\}, \{C_6, C_{10}\}, \{C_7, C_8, C_9\}$
10,000	$\{C_1, C_2, C_3\}, \{C_4, C_5\}, \{C_6, C_7, C_8\}, \{C_9, C_{10}\}$
100,000	$\{C_1, C_2, C_3\}, \{C_4, C_5\}, \{C_6, C_7, C_8\}, \{C_9, C_{10}\}$

Table 4.2: Performance of the algorithm on 3 simulated data with total counts of 1000, 10000, and 10000 in each diseased and control cases generated by (3.1)-(3.4) and parameters in Table 4.1

	0.57	0.74	0.56	0.67	0.00	0.00	0.00	0.00	0.00
0.57		0.61	0.73	0.67	0.00	0.00	0.00	0.00	0.00
0.74	0.61		0.59	0.68	0.00	0.00	0.00	0.00	0.00
0.56	0.73	0.59		0.66	0.00	0.00	0.00	0.00	0.00
0.67	0.67	0.68	0.66		0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00		0.54	0.52	0.42	0.76
0.00	0.00	0.00	0.00	0.00	0.54		0.68	0.67	0.54
0.00	0.00	0.00	0.00	0.00	0.52	0.68		0.70	0.53
0.00	0.00	0.00	0.00	0.00	0.42	0.67	0.70		0.45
0.00	0.00	0.00	0.00	0.00	0.76	0.54	0.53	0.45	

Table 4.3: A symmetric matrix recording the proportion of times alleles i and j appear together among all samples generated by the algorithm. Diagonal entries are all ones. Total counts is 1000 in both diseased and control cases

	0.95	0.93	0.16	0.08	0.00	0.00	0.00	0.00	0.00
0.95		0.94	0.15	0.07	0.00	0.00	0.00	0.00	0.00
0.93	0.94		0.17	0.09	0.00	0.00	0.00	0.00	0.00
0.00	0.15	0.17		0.89	0.00	0.00	0.00	0.00	0.00
0.16	0.07	0.09	0.89		0.00	0.00	0.00	0.00	0.00
0.08	0.00	0.00	0.00	0.00		0.96	0.53	0.03	0.00
0.00	0.00	0.00	0.00	0.00	0.96		0.54	0.05	0.01
0.00	0.00	0.00	0.00	0.00	0.53	0.54		0.47	0.37
0.00	0.00	0.00	0.00	0.00	0.03	0.05	0.47		0.85
0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.37	0.85	

Table 4.4: A symmetric matrix recording the proportion of times alleles i and j appear together among all samples generated by the algorithm. Diagonal entries are all ones. Total counts is 10000 in both diseased and control cases

	0.99	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.99		0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.99	0.99		0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.99		0.99	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00		0.99	0.99	0.03	0.00
0.00	0.00	0.00	0.00	0.00	0.99		0.99	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.99	0.99		0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00		1.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	

Table 4.5: A symmetric matrix recording the proportion of times alleles i and j appear together among all samples generated by the algorithm. Diagona entries are all ones. Total counts is 100000 in both diseased and control cases

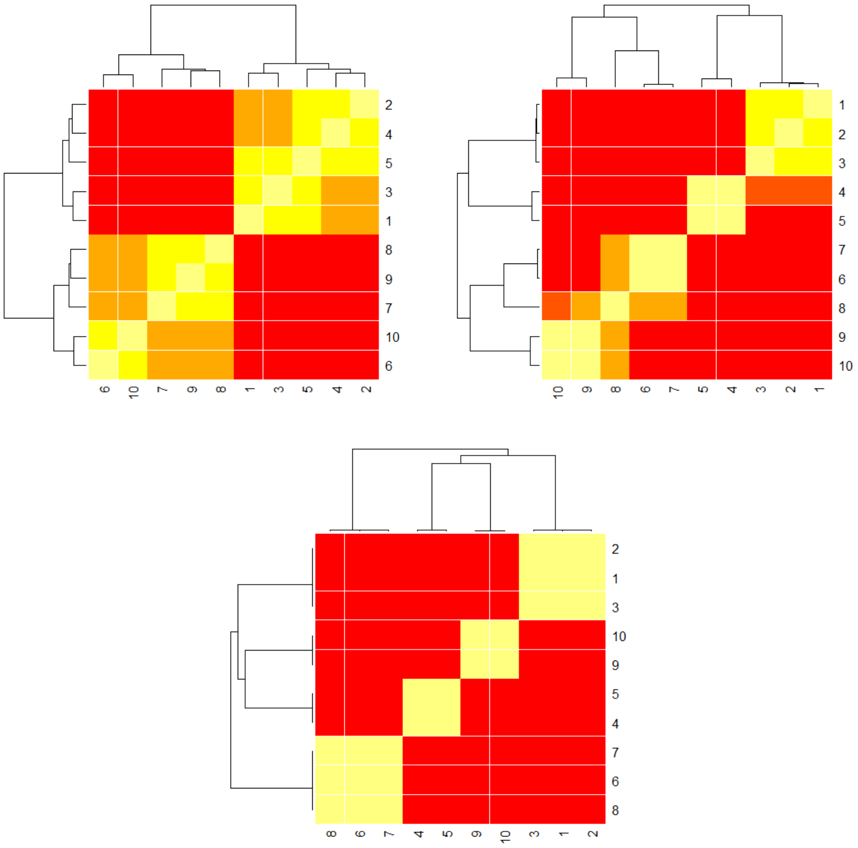


Figure 4.1: Heat maps capturing the grouping structure of the generated data with counts of 1000, 10000, and 100000.

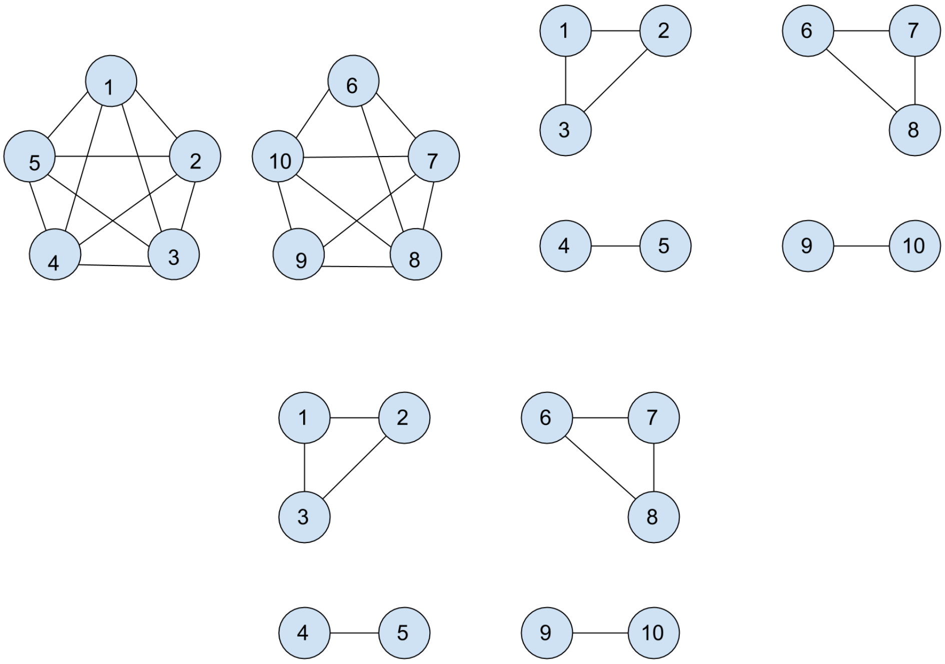


Figure 4.2: Groupings constructed from Tables 4.3 - 4.5 . Sample sizes are 1000 (left), 10000 (right), and 100000 (bottom). Cutoff = 0.5

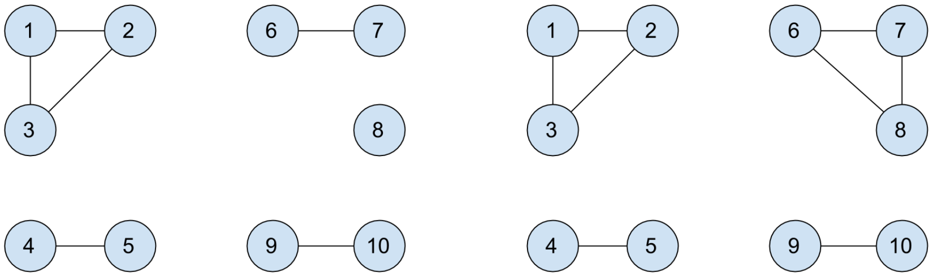


Figure 4.3: Groupings constructed from Tables 4.4 - 4.5 . Sample sizes are 10000 (left) and 100000 (right). Cutoff = 0.8

4.2 Analysis on 7 representative interactions found in RA

Now we are ready to apply the algorithm to the 7 interactions shown in Tables 2.1-2.2. The Metropolis Hastings algorithm (followed with local search algorithm) is run with 100,000 iterations. The first 1,000 generated samples are removed to allow for the burn-in period. As in the analysis of simulated data we refer the combination of Metropolis-Hastings and local search as the MAP algorithm.

Conversion table is provided for every interaction. This is needed since alleles with zero counts for both diseased and control pools are removed before the algorithm is run and hence the number $i \in \{1, ..., 27\}$ would not necessarily correspond to the same configuration in two different interactions. For each interaction, the MAP algorithm is run multiple times, with different initial values. Stable result is obtained for each interaction.

For each interaction, the grouping with maximal posterior is reported. It is then compared to groupings constructed by computing the proportion of time any two alleles i and j appear together. The reconstruction of the grouping is done twice with cutoff points of 0.25 and 0.5. These lower cutoff points were chosen as the probability of any pair of alleles appear together is generally smaller than the probability recorded from the generated data. Many entries of this proportion matrix are very close to zero. Each entry in the matrix is then mapped into a set of 5 colors. Entries with high probabilities are mapped to colors with low intensity (white) while entries with low probabilities are mapped to colors with high intensiy (red). We notice that a heat map is just another way to visualize the information contained in the two constructed groupings. Comparing the MAP estimate and the two reconstructed groupings, we notice that some parts of the grouping structure are captured by its reconstructions.

Here is the summary of group counts for grouping structures estimated by MAP algorithm, 0.25 and 0.5 cutoffs. The group counts for 0.25 and 0.5 cutoffs are determined by counting the number of connected components.

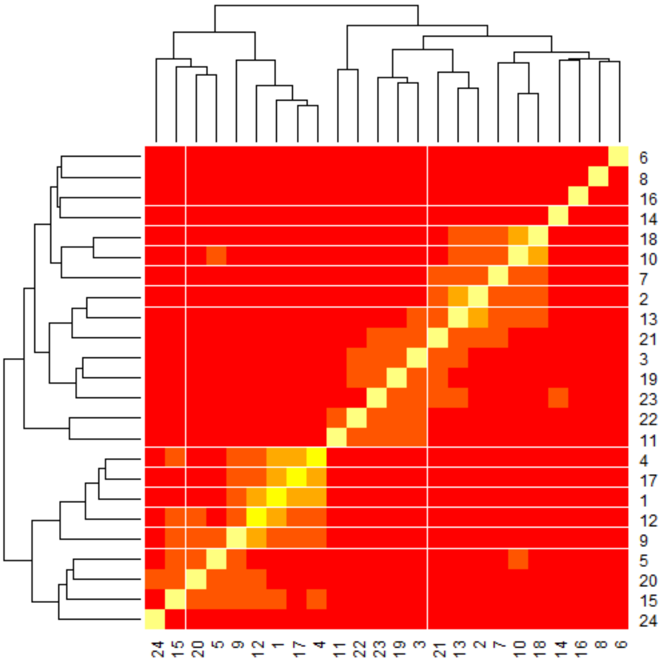
interaction	MAP	0.25 cutoff	0.5 cutoff
26	6	3	2
31	6	3	4
52	5	3	2
180	4	2	3
255	5	3	4
303	5	3	4
307	5	3	2

4.2.1 Interaction 26

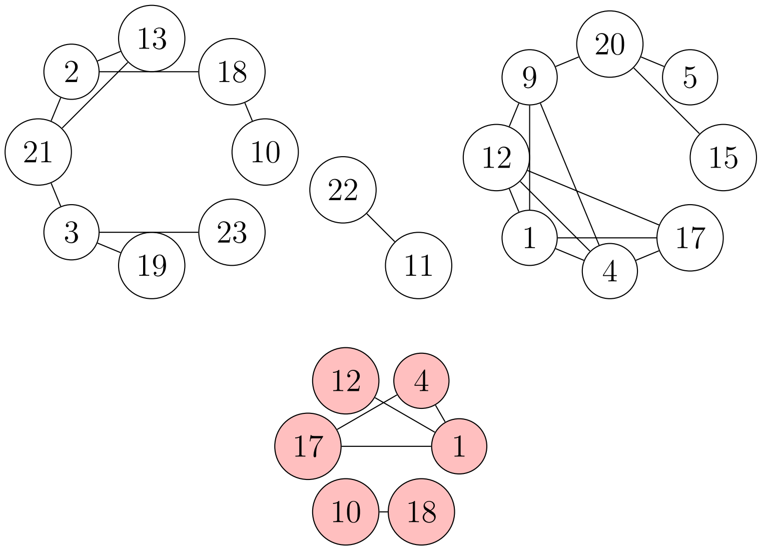
Conversion table for alleles in interaction 26

1	2	3	4	5	6	7	8	9	10	11	12
000	001	002	010	011	012	020	021	100	101	102	110
13	14	15	16	17	18	19	20	21	22	23	24
111	112	120	121	200	201	202	210	211	212	220	221

MAP estimate for int 26
$\{C_6, C_{10}, C_{16}, C_{18}, C_{24}\}$
$\{C_3, C_{14}, C_{19}, C_{23}\}$
$\{C_5, C_9, C_{15}, C_{20}\}$
$\{C_1, C_4, C_{12}, C_{17}\}$
$\{C_2, C_7, C_8, C_{13}, C_{21}\}$
$\{C_{11}, C_{22}\}$



Groupings with 0.25 (white) and 0.5 (pink) cutoffs

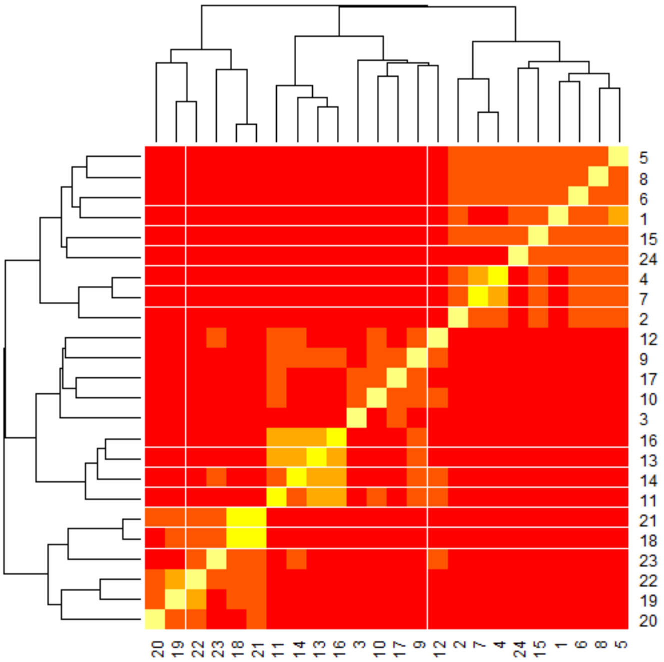


4.2.2 Interaction 31

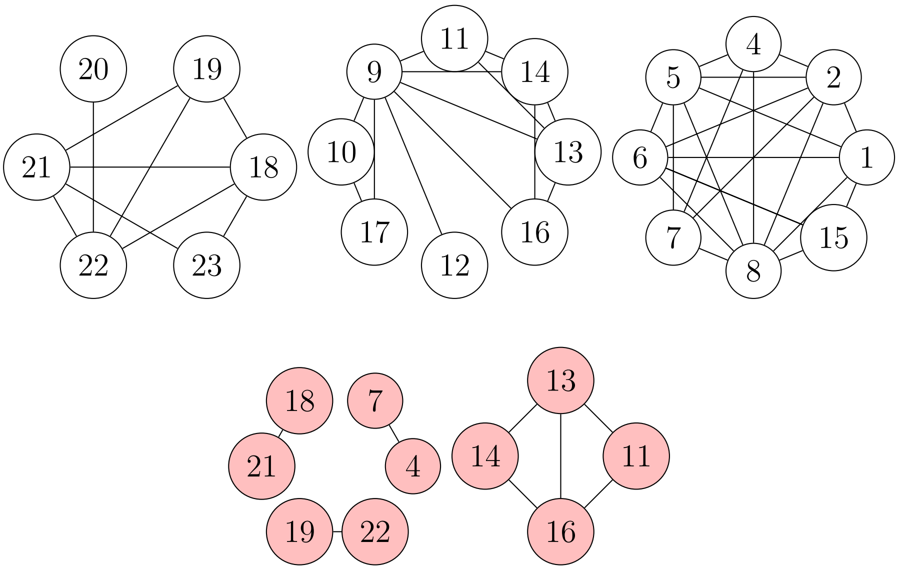
Conversion table for alleles in interaction 31

1	2	3	4	5	6	7	8	9	10	11	12
000	001	002	010	011	012	020	021	022	100	101	102
13	14	15	16	17	18	19	20	21	22	23	24
110	111	112	120	121	200	201	202	210	211	220	221

MAP estimate for int 31
$\{C_{18}, C_{21}, C_{23}\}$
$\{C_3, C_9, C_{10}, C_{12}, C_{17}\}$
$\{C_1, C_5, C_6, C_8, C_{15}, C_{24}\}$
$\{C_{11}, C_{13}, C_{14}, C_{16}\}$
$\{C_{19}, C_{20}, C_{22}\}$
$\{C_2, C_4, C_7\}$



Groupings with 0.25 (white) and 0.5 (pink) cutoffs



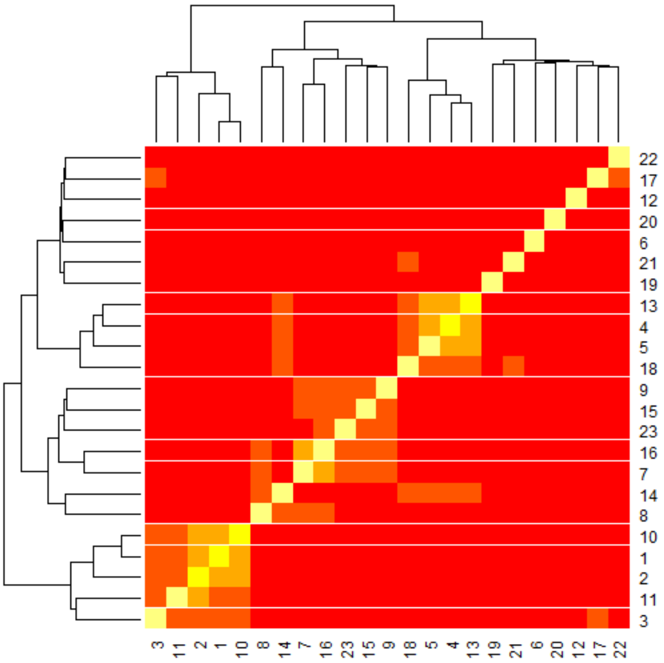
4.2.3 Interaction 52

Conversion table for alleles in interaction 52

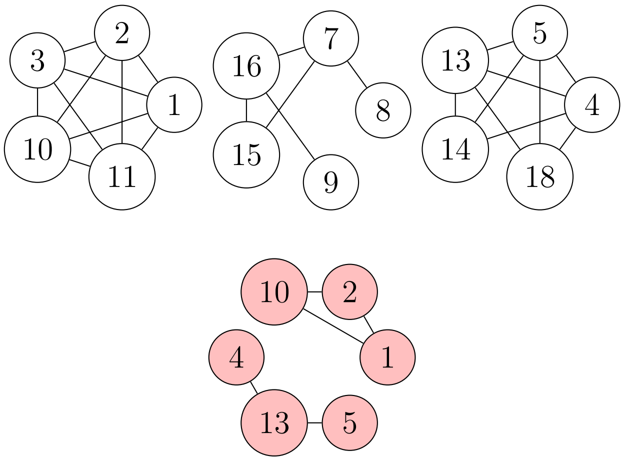
1	2	3	4	5	6	7	8	9	10	11	12
000	001	002	010	011	012	020	021	022	100	101	102

13	14	15	16	17	18	19	20	21	22	23
110	111	112	120	121	200	201	202	210	211	212

MAP estimate for int 52
$\{C_{12}, C_{19}, C_{20}, C_{22}\}$
$\{C_9, C_{15}, C_{23}\}$
$\{C_6, C_7, C_8, C_{16}\}$
$\{C_4, C_5, C_{13}, C_{14}, C_{18}, C_{21}\}$
$\{C_1, C_2, C_3, C_{10}, C_{11}, C_{17}\}$



Groupings with 0.25 (white) and 0.5 (pink) cutoffs

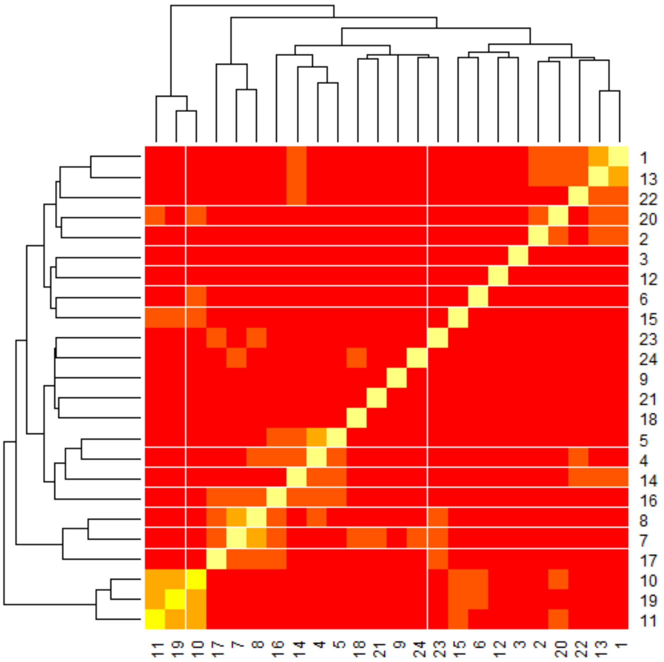


4.2.4 Interaction 180

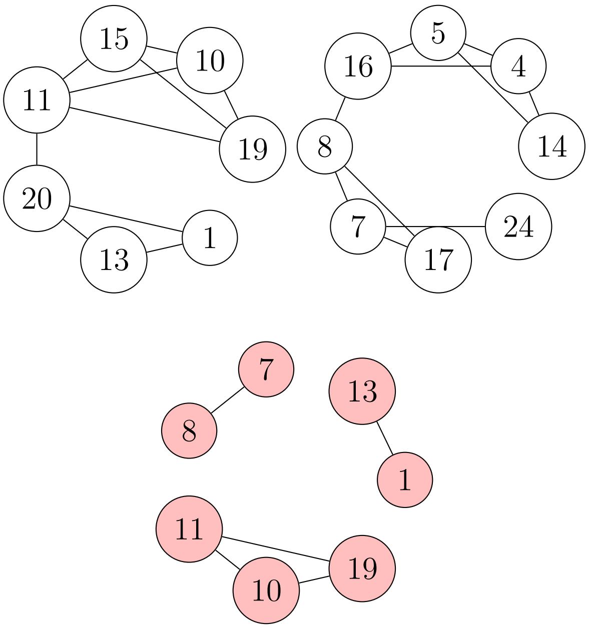
Conversion table for alleles in interaction 180

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
000	001	002	010	011	012	020	021	022	100	101	102	110	111	112	120	121	122	200	201	202	210	211	220

MAP estimate for int 180
$\{C_6, C_{10}, C_{11}, C_{15}, C_{19}\}$
$\{C_1, C_2, C_3, C_{13}, C_{20}, C_{22}\}$
$\{C_7, C_8, C_9, C_{17}, C_{18}, C_{21}, C_{23}, C_{24}\}$
$\{C_4, C_5, C_{12}, C_{14}, C_{16}\}$



Groupings with 0.25 (white) and 0.5 (pink) cutoffs



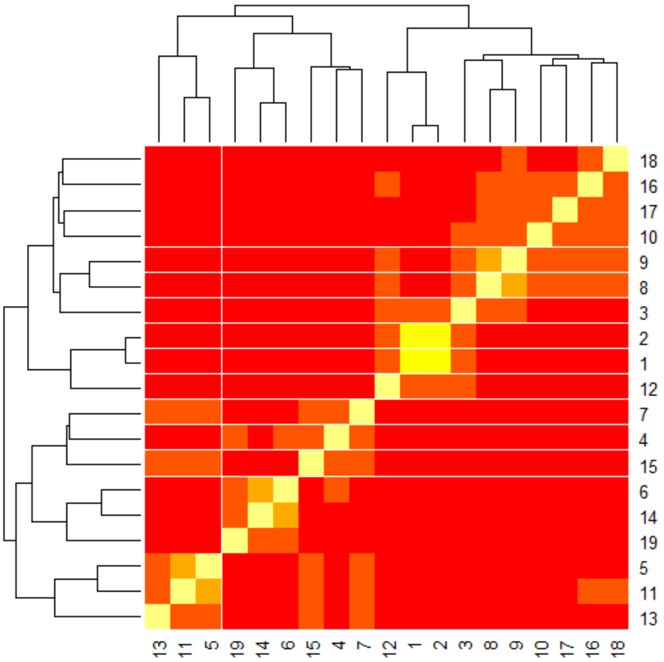
4.2.5 Interaction 255

Conversion table for alleles in interaction 255

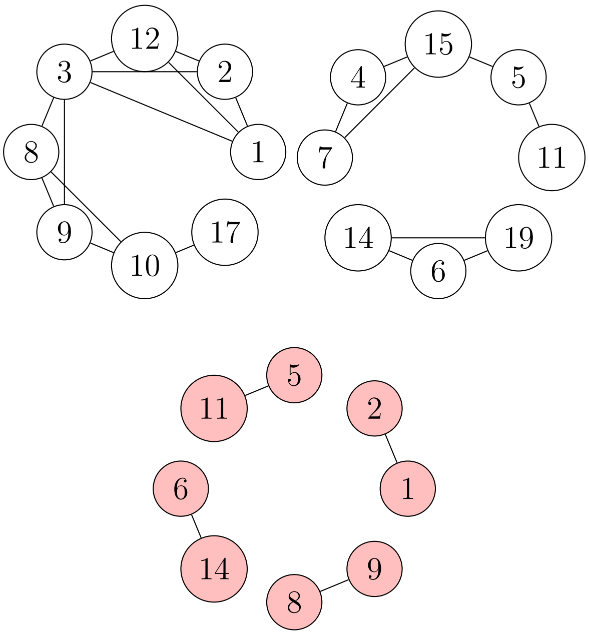
1	2	3	4	5	6	7	8	9	10
000	001	002	010	011	012	020	100	101	102

11	12	13	14	15	16	17	18	19
110	111	112	120	121	210	211	220	221

MAP estimate for int 255
$\{C_8, C_9, C_{10}, C_{16}, C_{17}\}$
$\{C_1, C_2, C_3, C_{12}\}$
$\{C_6, C_{14}\}$
$\{C_4, C_7, C_{15}, C_{19}\}$
$\{C_5, C_{11}, C_{13}, C_{18}\}$



Groupings with 0.25 (white) and 0.5 (pink) cutoffs

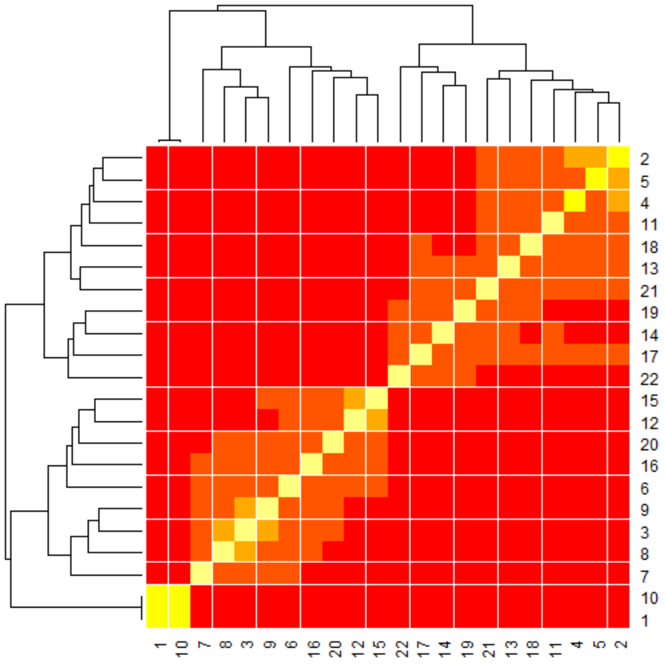


4.2.6 Interaction 303

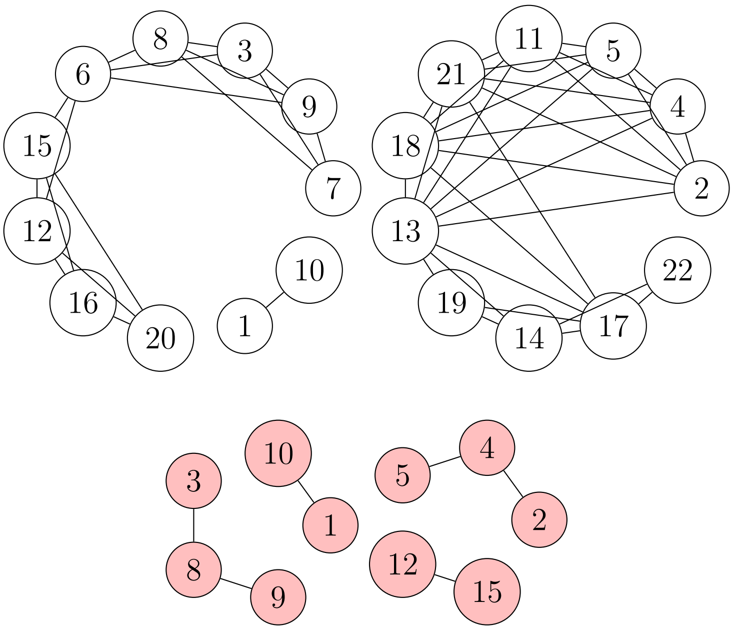
Conversion table for alleles in interaction 303

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
000	001	002	010	011	012	020	021	022	100	101	102	110	111	112	120	121	200	201	202	210	211

MAP estimate for int 303
$\{C_2, C_4, C_5, C_{11}, C_{13}, C_{18}, C_{21}\}$
$\{C_3, C_6, C_7, C_8, C_9\}$
$\{C_{14}, C_{17}, C_{19}, C_{22}\}$
$\{C_{12}, C_{15}, C_{16}, C_{20}\}$
$\{C_1, C_{10}\}$



Groupings with 0.25 (white) and 0.5 (pink) cutoffs

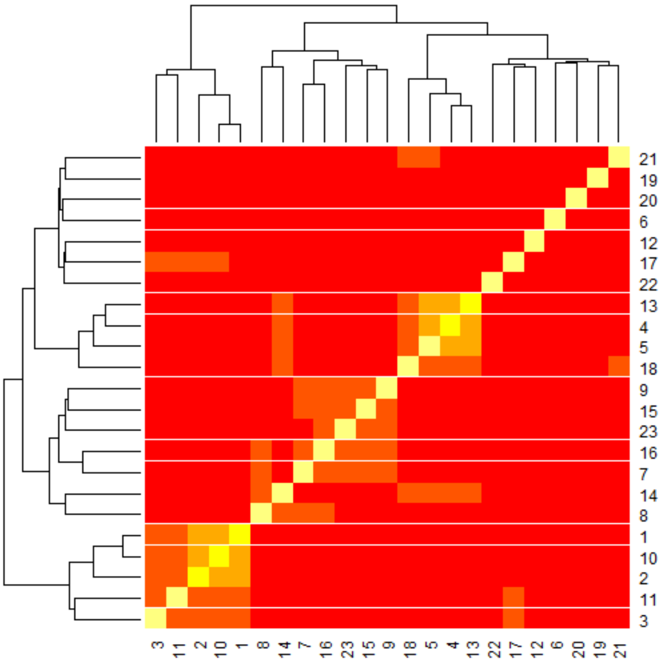


4.2.7 Interaction 307

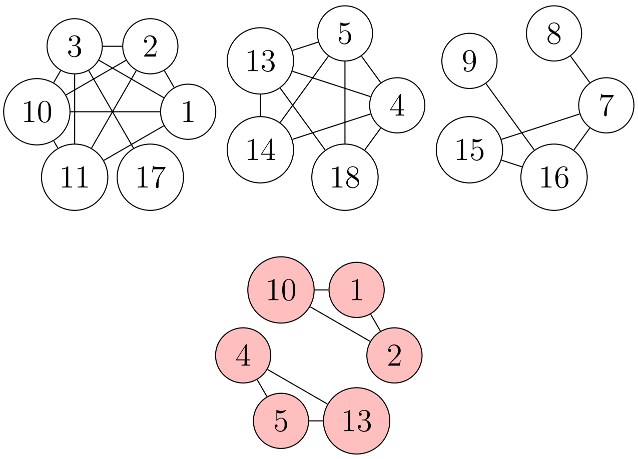
Conversion table for alleles in interaction 307

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
000	001	002	010	011	012	020	021	022	100	101	102	110	111	112	120	121	200	201	202	210	211	212

MAP estimate for int 307
$\{C_6, C_9, C_{20}, C_{22}, C_{23}\}$
$\{C_1, C_2, C_{10}, C_{11}\}$
$\{C_3, C_{12}, C_{17}\}$
$\{C_4, C_5, C_{13}, C_{14}, C_{18}, C_{21}\}$
$\{C_7, C_8, C_9, C_{15}, C_{16}\}$



Groupings with 0.25 (white) and 0.5 (pink) cutoffs



Let us take a closer look at interaction 307. Results for other interactions in this thesis could be studied similarly. The MAP estimate and 0.25 cutoff show similar grouping structure. However, $\{C_1, C_2, C_{10}, C_{11}\}$ and $\{C_3, C_{17}\}$ found by MAP are collapsed into one big group in 0.25 cutoff. The two cutoff points allow us to compare the strength of the grouping structure. Some edges, and hence alleles, are eliminated as stricter cutoff point of 0.5 is chosen. The very same information could be read from the heat map. We notice lighter blocks $\{C_1, C_2, C_{10}\}$ and $\{C_4, C_5, C_{13}\}$ are located within blocks $\{C_1, C_2, C_3, C_{10}, C_{11}\}$ and $\{C_4, C_5, C_{13}, C_{14}, C_{18}\}$, respectively. Alleles that are not included in constructed groupings are shown as individual blocks. We also observe a few cliques in the connected components of the grouping with 0.25 cutoff. A clique in a graph is a subcollection of vertices such that any two vertices are connected by an edge. These cliques suggest that some alleles in one group might be more connected to each other than the rest of the alleles in the same group. We notice that $\{C_1, C_2, C_3, C_{10}, C_{11}\}$ forms a clique in $\{C_1, C_2, C_3, C_{10}, C_{11}, C_{17}\}$ whereas C_{17} is connected to the rest of the alleles only through C_3.

We might try to dramatically increase the number of iterations of the Metropolis-Hastings algorithm, say to 1000,000, in order to check whether we could improve the estimate of the grouping stucture for each interaction. This is true since, assuming that the Markov Chain would not be trapped in a local mode, we let the constructed Markov Chain to further explore the states and hope that a state with high posterior probability would be visited. This might take more time but such experiment is appropriate whenever one has efficient programming facilities. We might also try to assign a value other than one to γ. A different grouping stucture might have been found if we were to do this. A value of γ different from one would induce a change in the computation of the posterior distribution of the grouping, assigning higher probabilities to the groupings with smaller L.

CHAPTER 5

Conclusion

The MAP algorithm performs well when it is run on generated data with a sufficiently large total counts. We saw that the algorithm finds the correct grouping $(S_1, S_2, S_3, S_4) = (\{C_1, C_2, C_3\}, \{C_4, C_5\}, \{C_6, C_7, C_8\}, \{C_9, C_{10}\})$ as it is applied to the data with total counts of 10,000 in each diseased and control pools. The same thing is true when total counts of both pools are increased to 100,000. The relatively small total counts combined with the randomness in the generation of the first data set result in a slightly inaccurate identification of the true grouping. This defect might be remedied by increasing the number of iterations in the Metropolis-Hastings algorithm. In the case of the data set with 100,000 total counts, the MAP estimate is identical to the reconstruction of the grouping even when a very strict cutof point of 0.95 is chosen.

Stable results are obtained for every interactions found in RA. However, a more lenient cutoff points of 0.25 and 0.5 are chosen to reconstruct the grouping for each interaction. Entries in the symmetric proportion matrix are generally smaller than those of the matrix for simulated data. Even though the grouping found by the MAP algorithm is not identical to the reconstructed grouping in every interaction, the reconstructed grouping still captures partial structure of the MAP estimate. These discrepancies might have been caused by the small counts (less that 1000 patients in both diseased and control pools) in every interaction. Moreover, it seems that the probability distribution of the allele counts widely vary in every interaction studied. For instance, we observe 763 diseased counts and 497 control counts for allele C_1 but many small counts, like 0, 1, and 2, are observed in both diseased and control pools for many other alleles in interaction 255. Such might be troublesome since randomness and small probabilities in the assumed model for data generation could easily disrupt the ratio of diseased and control counts. It is also possible that the actual data generation in the case of the Rheumatoid Arthritis slightly departs from our model of data generation.

CHAPTER 6

Appendix: Mathematical Derivation

In this section, we show the full derivation of the likelihood function shown in (3.8). We know that:

$$
\begin{aligned}
&P(n_1, ..., n_m, n'_1, ..., n'_m | S_1, ..., S_L) \\
&= \int_{\vec{\theta}_1,...,\vec{\theta}_L} P(n_1, ..., n_m, n'_1, ..., n'_m, \vec{\theta}_1, ..., \vec{\theta}_L | S_1, ..., S_L) d(\vec{\theta}_1, ..., \vec{\theta}_L) \\
&= \int_{\vec{\theta}_1,...,\vec{\theta}_L} P(n_1, ..., n_m, \vec{\theta}_1, ..., \vec{\theta}_L | S_1, ..., S_L) P(n'_1, ..., n'_m, \vec{\theta}_1, ..., \vec{\theta}_L | S_1, ..., S_L) \\
&\qquad P(\vec{\theta}_1, ..., \vec{\theta}_L | S_1, ..., S_L) d(\vec{\theta}_1, ..., \vec{\theta}_L)
\end{aligned}
\tag{6.1}
$$

where the second equality follows from conditional independence between $n_1, ..., n_m$, and $n'_1, ..., n'_m$ given $\vec{\theta}_1, ..., \vec{\theta}_L, S_1, ..., S_L$. We are going to consider each term separately.

In order to do this, we first need to list several assumptions and consequences of the model described in chapter 3:

- $P(n_1, ..., n_m, | \vec{p}, \vec{\theta}_1, ..., \vec{\theta}_L, S_1, ..., S_L) = \prod_{i=1}^{L} \left(p_i^{n_{S_i}} \times \prod_{j=1}^{|S_i|} \theta_{ij}^{n_{ij}} \right)$

  We note that this is a consequence of (3.1) - (3.4). It simply says that the probability of observing $n_1, ..., n_m$ given the model and a fixed partition $S_1, ..., S_L$ is determined by the allele counts within and between the groups.

- $P(\vec{\theta}_1, ..., \vec{\theta}_L | S_1, ..., S_L) = \prod_{i=1}^{L} P(\vec{\theta}_i | S_i)$

  We assume that, conditional on the grouping $S_1, ..., S_L$, the probability distribution of allele counts within one group is independent from the probability distribution of allele counts in different group.

- $P(\vec{p}|\vec{\theta}_1, ..., \vec{\theta}_L, S_1, ..., S_L) = P(\vec{p}|L)$

  Given that information about L is known, we don't need the knowwldge of the exact grouping nor the probability distributions of allele counts in each group.

We first compute the first term in the equation (6.1). The second term is computed the same way. Recall that $P(\vec{p}|L)$ and $P(\vec{p}|L) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_L)$. By integrating over $\vec{p}$ (or over $\vec{p}'$ for the second term) , we have:

$$\begin{aligned}
&P(n_1, ..., n_m, \vec{\theta}_1, ..., \vec{\theta}_L|S_1, ..., S_L) \\
&= \int_{\vec{p}} P(n_1, ..., n_m, |\vec{p}, \vec{\theta}_1, ..., \vec{\theta}_L, S_1, ..., S_L) P(\vec{p}|\vec{\theta}_1, ..., \vec{\theta}_L, S_1, ..., S_L) d\vec{p} \\
&= \int_{\vec{p}} \prod_{i=1}^{L} \left(p_i^{n_{S_i}} \times \prod_{j=1}^{|S_i|} \theta_{ij}^{n_{ij}} \right) \times P(\vec{p}|L) d\vec{p} \\
&= \int_{\vec{p}} \prod_{i=1}^{L} \left(p_i^{n_{S_i}} \times \prod_{j=1}^{|S_i|} \theta_{ij}^{n_{ij}} \right) \times \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\prod_{i=1}^{L} \Gamma(\alpha_i)} \times \prod_{i=1}^{L} p_i^{\alpha_i - 1} d\vec{p} \qquad (6.2) \\
&= \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\prod_{i=1}^{L} \Gamma(\alpha_i)} \times \prod_{i=1}^{L} \left(\prod_{j=1}^{|S_i|} \theta_{ij}^{n_{ij}} \right) \times \int_{\vec{p}} \prod_{i=1}^{L} p_i^{n_{S_i} + \alpha_i - 1} d\vec{p} \\
&= \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\Gamma(\sum_{i=1}^{L} n_{S_i} + \alpha_i)} \times \prod_{i=1}^{L} \left(\frac{\Gamma(n_{S_i} + \alpha_i)}{\Gamma(\alpha_i)} \right) \times \prod_{i=1}^{L} \left(\prod_{j=1}^{|S_i|} \theta_{ij}^{n_{ij}} \right)
\end{aligned}$$

Incorporating $P(\vec{\theta}_1, ..., \vec{\theta}_L|S_1, ..., S_L) = \prod_{i=1}^{L} P(\vec{\theta}_i|S_i)$, we are ready to finish the derivation of the likelihood function.

$$
\begin{aligned}
&P(n_1, ..., n_m, n'_1, ..., n'_m|S_1, ..., S_L)\\
&= \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\Gamma(\sum_{i=1}^{L} n_{S_i} + \alpha_i)} \times \prod_{i=1}^{L} \frac{\Gamma(n_{S_i} + \alpha_i)}{\Gamma(\alpha_i)} \times \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\Gamma(\sum_{i=1}^{L} n'_{S_i} + \alpha_i)} \times \prod_{i=1}^{L} \frac{\Gamma(n'_{S_i} + \alpha_i)}{\Gamma(\alpha_i)}\\
&\quad \int_{\vec{\theta}} \prod_{i=1}^{L} \left(\prod_{j=1}^{|S_i|} \theta_{ij}^{n_{ij}+n'_{ij}} \right) \times \prod_{i=1}^{L} \left(\frac{\Gamma(\sum_{j=1}^{|S_i|} \frac{1}{|S_i|})}{\prod_{j=1}^{|S_i|} \Gamma(\frac{1}{|S_i|})} \prod_{j=1}^{|S_i|} \theta_{ij}^{\frac{1}{|S_i|}-1} \right) d\vec{\theta}\\
&= \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\Gamma(\sum_{i=1}^{L} n_{S_i} + \alpha_i)} \times \prod_{i=1}^{L} \frac{\Gamma(n_{S_i} + \alpha_i)}{\Gamma(\alpha_i)} \times \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\Gamma(\sum_{i=1}^{L} n'_{S_i} + \alpha_i)} \times \prod_{i=1}^{L} \frac{\Gamma(n'_{S_i} + \alpha_i)}{\Gamma(\alpha_i)}\\
&\quad \prod_{i=1}^{L} \left(\frac{\Gamma(\sum_{j=1}^{|S_i|} \frac{1}{|S_i|})}{\prod_{i=1}^{|S_i|} \Gamma(\frac{1}{|S_i|})} \right) \times \int_{\vec{\theta}} \prod_{i=1}^{L} \left(\prod_{j=1}^{|S_i|} \theta_{ij}^{n_{ij}+n'_{ij}+\frac{1}{|S_i|}-1} \right) d\vec{\theta}\\
&= \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\Gamma(\sum_{i=1}^{L} n_{S_i} + \alpha_i)} \times \prod_{i=1}^{L} \frac{\Gamma(n_{S_i} + \alpha_i)}{\Gamma(\alpha_i)} \times \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\Gamma(\sum_{i=1}^{L} n'_{S_i} + \alpha_i)} \times \prod_{i=1}^{L} \frac{\Gamma(n'_{S_i} + \alpha_i)}{\Gamma(\alpha_i)}\\
&\quad \prod_{i=1}^{L} \left(\frac{\Gamma(\sum_{j=1}^{|S_i|} \frac{1}{|S_i|})}{\prod_{i=1}^{|S_i|} \Gamma(\frac{1}{|S_i|})} \right) \times \prod_{i=1}^{L} \int_{\vec{\theta}_i} \left(\prod_{j=1}^{|S_i|} \theta_{ij}^{n_{ij}+n'_{ij}+\frac{1}{|S_i|}-1} \right) d\vec{\theta}_i \text{ by Fubini}\\
&= \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\Gamma(\sum_{i=1}^{L} n_{S_i} + \alpha_i)} \times \prod_{i=1}^{L} \frac{\Gamma(n_{S_i} + \alpha_i)}{\Gamma(\alpha_i)} \times \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\Gamma(\sum_{i=1}^{L} n'_{S_i} + \alpha_i)} \times \prod_{i=1}^{L} \frac{\Gamma(n'_{S_i} + \alpha_i)}{\Gamma(\alpha_i)}\\
&\quad \prod_{i=1}^{L} \frac{\Gamma(\sum_{j}^{|S_i|} \frac{1}{|S_i|})}{\Gamma(\sum_{j}^{|S_i|} n_{ij} + n'_{ij} + \frac{1}{|S_i|})} \times \prod_{i=1}^{L} \left(\prod_{j=1}^{|S_i|} \frac{\Gamma(n_{ij} + n'_{ij} + \frac{1}{|S_i|})}{\Gamma(\frac{1}{|S_i|})} \right)\\
&= \frac{1}{\Gamma(1+n)} \prod_{i=1}^{L} \frac{\Gamma(n_{S_i} + \frac{1}{L})}{\Gamma(\frac{1}{L})} \times \frac{1}{\Gamma(1+n')} \prod_{i=1}^{L} \frac{\Gamma(n'_{S_i} + \frac{1}{L})}{\Gamma(\frac{1}{L})}\\
&\quad \prod_{i=1}^{L} \frac{1}{\Gamma(1 + n_{S_i} + n'_{S_i})} \times \prod_{i=1}^{L} \prod_{j=1}^{|S_i|} \frac{\Gamma(n_{ij} + n'_{ij} + \frac{1}{|S_i|})}{\Gamma(\frac{1}{|S_i|})}
\end{aligned}
\tag{6.3}
$$

where the last equality is obtained by letting $\alpha_i = \frac{1}{L}$, i.e. we assign uniform prior for $\vec{\theta}_i, \vec{p}$, and $\vec{p}'$.

References

[1] Cornelis, F., S. Faure, M. Martinez, J.F. Prud'homme and P. Fritz et al., 1998. *New susceptibility locus for rheumatoid arthritis suggested by a genome-wide linkage study* Proc. Nat. Acad. Sci., 95:10746-10750.

[2] Manolio, T.A., F.S. Collins, N.J. Cox, D.B. Goldstein and L.A. Hindroff et al., 2009. *Finding the missing heritability of complex diseases* Nature, 461:747-753.

[3] Plenge, R.M., M. Seiselstad, L Padyukov, A.T. Lee and E.F. Remmers, 2007. *TRAFI-C5 as a risk locus for rheumatoid arthritis–a genomewide study* N. Engl. J, Med., 357;1199-1209.

[4] Stahl, E.A., S. Raychaudhuri, E.F. Remmers, G.Xie and S. Eyre et al. 2010. *Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci* Nat. Gen., 42:508-514.

[5] Wu, Z and H. Zhao, 2009. *Statistical power of model selection strategies for genome-wide association studies* PLoS Genet., 5; 849-911.

[6] Zhang, J., Z. Wu, C. Gao and M.Q. Zhang, 2012. *High-order interactions in rheumatoid arthritis detected by Bayesian method using genome-wide association studies data.* Am. Med. J., 3: 56-66.

[7] Zhang, Y. and J.S. Liu, 2007. *Bayesian inference of epistatic interactions in case-control studies.* Nat. Genet., 39:1167-1173

[8] Zhang, Y., J. Zhang and J.S. Liu, 2011 *Block-based Bayesian epistasis association mapping with application to WTCCC type 1 diabetes data* Ann. Applied Stat., 5: 2052-2077