entropy

MDPI

Article

Bayesian Genetic Association Test when Secondary Phenotypes Are Available Only in the Case Group

Yongku Kim [1,†] and Minjung Kwak [2,*,†]

1 Department of Statistics, Kyungpook National University, Daegu 41566, Korea; kim.1252@knu.ac.kr
2 Department of Statistics, Yeungnam University, Gyeongsan 38541, Korea
* Correspondence: mjkwak@yu.ac.kr; Tel.: +82-53-810-2321; Fax: +82-53-810-4615
† These authors contributed equally to this work.

Academic Editor: Raúl Alcaraz Martínez
Received: 14 September 2015; Accepted: 7 December 2015; Published: 6 April 2016

Abstract: In many case-control genetic association studies, a secondary phenotype that may have common genetic factors with disease status can be identified. When information on the secondary phenotype is available only for the case group due to cost and different data sources, a fitting linear regression model ignoring supplementary phenotype data may provide limited knowledge regarding genetic association. We set up a joint model and use a Bayesian framework to estimate and test the effect of genetic covariates on disease status considering the secondary phenotype as an instrumental variable. The application of our proposed procedure is demonstrated through the rheumatoid arthritis data provided by the 16th Genetic Analysis Workshop.

Keywords: genome-wide association study; Bayesian testing; secondary phenotype; Bayes factor; case-control design; incomplete data

1. Introduction

Statistical analysis with regression models is a common approach for investigating the effects of genetic and non-genetic covariates on a response outcome. Observations in such a study usually involve a pre-specified disease trait, which could be either a continuous or categorical variable, genetic markers, such as single nucleotide polymorphisms (SNPs), candidate genes and non-genetic covariates of interest, such as age, ethnicity and environmental exposure. When the disease trait is obtained as a continuous outcome variable, linear regression models are commonly used [1,2], whereas logistic regression models, the Cochran–Armitage trend test and Pearson's chi-squared test are often used for the analysis of a binary disease trait [3,4]. In addition, the effects of genes (SNPs) and covariates on the selected disease trait are often described through a multivariate linear or generalized linear model. Through a set of latent biological phenotypes, Chatterjee *et al.* [5] described a conceptual framework for modeling genetic associations and gene-gene and gene-environment interactions in indirect-association studies with multivariate logistic regression models. Maity *et al.* [6] extended the approaches proposed by Chatterjee *et al.* [5] to studies with repeated measures data and developed a class of score tests in general semi-parametric regression models. Genetic studies may also involve combined data from several case-control studies [7].

Recently, Wu *et al.* [8] proposed a joint regression model based on a two-step conditional modeling approach. First, they modeled the effects of SNPs and non-genetic covariates on the conditional probabilities for the categorical variable using a generalized linear model; second, based on conditioning of the given levels of a categorical variable, they modeled the added effects of the same SNPs and covariates on the conditional means of the continuous outcome using a general linear regression model. However, they modeled the conditional mean of the continuous outcome for given case and control groups separately; thus, their joint likelihood would not allow for the

estimation of the effect of covariates on the continuous outcome measure in the control group when the continuous outcomes are missing in the control group. Furthermore, He *et al.* [9] proposed a Gaussian copula-based approach that could model the dependence between disease status and secondary phenotypes.

As progress in computational power has been extended to genetic association studies, recent studies have demonstrated the practical and theoretical advantages of using Bayesian approaches for the assessment of associations [10]. Bayesian methods compute measures of evidence that can be directly compared among SNPs within and across studies. In genome-wide association studies (GWAS), the Bayes factor is a useful tool to support significant *p*-values and serves as a better measure than p-values when results are compared across studies with different sample sizes. Xu *et al*. [11] proposed a Bayes factor based on the Cochran–Armitage trend test that incorporates situations of Hardy–Weinberg disequilibrium.

In this paper, we propose an instrumental regression model and develop a Bayesian estimation and testing procedure involving the shared genetic associations with both continuous and binary outcomes. We construct a joint-likelihood model for disease status and the continuous secondary phenotype available for the case group. In particular, we incorporate the secondary phenotype as an instrumental variable into the model and make use of the fact that the missing secondary phenotypes are known to be larger (or smaller) for the control group than for the case group. This is not an unreasonable assumption in practice, since the available phenotype is regarded as an important surrogate clinical measurement in deciding the disease status. Under various genetic models, estimation of the genetic associations and covariate effects are obtained by the Bayesian method.

For the main results, we describe the real data and data structure in Sections 2 and 3, respectively. We describe our model and the hypotheses to be tested in Section 4 and then explain the Bayesian methodology and the corresponding testing procedures in Section 5. In Section 6, we present the application of our method to the Genetic Analysis Workshop 16 (GAW16) rheumatoid arthritis (RA) dataset. In Section 7, we provide concluding remarks with a brief discussion of the advantages and limitations of our approach.

2. RA Genetic Data

Our example comes from a GWAS exploring the effect of genetic markers on certain diseases. GWAS has emerged as an effective tool to identify a common polymorphism underlying complex diseases, and the GAW16 RA data represent the initial batch of GWAS data from the North American Rheumatoid Arthritis Consortium after removing duplicated and contaminated samples [12].

RA is a chronic inflammatory disease characterized by the destruction of the synovial joints, resulting in severe disability, particularly in patients who remain refractory to available therapies. The susceptibility to and severity of RA are determined by both genetic and environmental factors [13]. A newly-identified autoantibody, anti-cyclic citrullinated peptide (anti-CCP), appears to be highly specific for the disease and is a good predictor of erosive outcome [14]; elevations of anti-CCP have been reported to predict increased risk for development of RA [15]. RA affection is a dichotomous variable representing disease status, and anti-CCP level is a continuous trait. Moreover, information of the gender of each patient is available as an independent covariate.

The GAW16 RA data are derived from a case-control study involving 868 RA-positive patients (cases) and 1194 subjects from the New York Cancer Project, who were RA negative (controls), and the dataset contains genotype data for more than 500,000 SNPs. Since anti-CCP antibodies are potentially important surrogate markers for the diagnosis and prognosis of RA and larger anti-CCP values have been linked to increased severity of RA [14], the GAW16 data contain anti-CCP measurements for the RA-positive patients, but not for the RA-negative subjects.

For the purpose of illustration, we here present the results for a few selected SNPs on chromosome 1. Among susceptibility genes on chromosome 1, PTPN22 has been shown to be

associated with RA [16], and PADI4 has been identified to be associated with RA among Asians [17]. Alleles at the PTPN22 locus have been shown to confer an increased risk for RA [18]. In particular, they reported that the R620W allele in SNP rs2476601 confers a 1.7-fold to 1.9-fold increased risk of RA to heterozygotes and even higher risks to homozygous carriers.

After removing the SNPs that were excluded using standard GWAS quality-control procedures, we obtained 34,195 SNPs on chromosome 1. Due to the skewness of the anti-CCP distribution, we used the log-transformed anti-CCP values. For preliminary analysis, we screened each individual SNP based on the results of logistic regression of RA status under an additive genetic model and selected SNPs whose p-values were smaller than $10^{-5.83}$ (Bonferroni-corrected significance level). In Table 1, we have tabulated the top 13 ranked SNPs that were screened from the preliminary logistic analysis.

Table 1. The 13 most significant single nucleotide polymorphisms (SNPs) from the logistic regression of rheumatoid arthritis status under the additive model.

Rank	SNP	p-value	MAF *
1	rs9442372	$10^{-15.95}$	0.430
2	rs2476601	$10^{-10.93}$	0.084
3	rs6427128	$10^{-7.70}$	0.129
4	rs2062629	$10^{-7.52}$	0.140
5	rs356116	$10^{-7.03}$	0.139
6	rs16861613	$10^{-6.93}$	0.074
7	rs6671416	$10^{-6.83}$	0.129
8	rs7524233	$10^{-6.50}$	0.141
9	rs6598886	$10^{-6.34}$	0.046
10	rs1046269	$10^{-6.20}$	0.115
11	rs11578154	$10^{-6.04}$	0.071
12	rs12027585	$10^{-6.02}$	0.113
13	rs2986742	$10^{-5.84}$	0.085

* MAF, minor allele frequency.

3. Data Structure

We consider the example of genetic association studies based on a case-control sample. Let $(D, W, \mathbf{G}, \mathbf{Z})$ be the random variables for the traits, genetic markers and covariates being considered, where D is a binary random variable indicating a certain disease or clinical status; W is the actual value of the secondary phenotype, $\mathbf{G} = (G_1, \ldots, G_q)^T$; $q \geq 1$ is a vector of categorical variables denoting the presence of genetic markers or SNPs; and $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$ is an $\mathbb{R}^p$-valued covariate matrix for some $p \geq 1$. Although $\mathbf{G}$ is generally a categorical variable, specific choices for $\mathbf{G}$ depend on whether the biallelic genetic model is recessive, additive or dominant. In practice, we identify a group of case subjects and select a certain number of control subjects matched with respect to other important characteristics, such as gender or age. Suppose that the data consist of n independent and identically distributed realizations of $(D, W, \mathbf{G}, \mathbf{Z})$, where the observation for the i-th case subject is denoted by $\{(1, W_i, \mathbf{G}_i, \mathbf{Z}_i),\ i = 1, \ldots, r\}$ and the observation for the j-th control subject is denoted by $\{(0, \mathbf{G}_i, \mathbf{Z}_i),\ i = r+1, \ldots, n\}$. Although D_is are observed for all subjects $i = 1, \ldots, n$, because of cost considerations or a situation in which the secondary phenotype W in the control group may not have an important impact on the scientific objectives of the study, W_i may not be measured in the control group. For instance, in our example, it is well known that high anti-CCP values are associated with RA, and these values were measured for the case subjects. However, control group data were obtained from a different source, *i.e.*, the New York Cancer Project, and therefore, the anti-CCP values are missing for the control subjects. On the other hand, the secondary phenotype serves as an important surrogate measurement for disease status; therefore, we may assume that there is a certain ordering in magnitude of the secondary phenotype values between the case group

and the control group. Throughout this paper, we assume known information that W's in the control group are smaller than those in the case group. The opposite situation can be easily dealt with by taking negative values of the secondary phenotype outcomes.

4. Model and Hypothesis

We consider two response outcomes, namely the binary disease status D and the continuous secondary phenotype outcome W. In a classical logistic regression setup, the probability of having a disease is determined by the covariates in the model. Here, we also evaluate the values of a continuous secondary phenotype, which are used as a surrogate measure to determine the disease status. The secondary phenotype W is deemed to be correlated with the disease status D and is also explained by the covariates $\mathbf{G}$ and $\mathbf{Z}$. Our approach relies on being able to identify a genetic variant (SNP) that perturbs the secondary phenotype, which is measured with a random error parameter. We assume that the genetic markers, as well as other non-genetic covariates, have some effect on the secondary phenotype, which serves as an intermediate outcome that, in turn, affects the disease status. For instance, because elevated anti-CCP antibody levels are associated with an increased risk of RA, we assume that this relationship may be causal. Therefore, we investigated the effect of covariates on RA status via their effect on the anti-CCP measurements as a secondary outcome. This leads to the path diagram illustrated as follows.

$$(\mathbf{G}, \mathbf{Z}) \quad \longrightarrow \quad W \quad \longrightarrow \quad D$$

[Association diagram among variables]

We consider a joint regression model for (D_i, W_i) with covariates $\{\mathbf{G}_i, \mathbf{Z}_i\}$ constructed by modeling the conditional probability $P(D_i = 1|W_i, \mathbf{G}_i, \mathbf{Z}_i)$ and the conditional mean of W_i, given $\{\mathbf{G}_i, \mathbf{Z}_i\}$. Using a logit model for the binary variable D_i and a linear model for the continuous secondary phenotype W_i, a generalized linear joint model is given as the following hierarchical model based on $P(\mathbf{D}|\mathbf{W}, \mathbf{G}, \mathbf{Z})$ and $P(\mathbf{W}|\mathbf{G}, \mathbf{Z})$:

$$\begin{cases} \text{logit}(P(D_i = 1|W_i, \mathbf{G}_i, \mathbf{Z}_i)) = \beta_0 + \boldsymbol{\beta}_1^T \mathbf{G}_i + \boldsymbol{\beta}_2^T \mathbf{Z}_i + \boldsymbol{\beta}_3^T W_i \\ W_i = \gamma_0 + \boldsymbol{\gamma}_1^T \mathbf{G}_i + \boldsymbol{\gamma}_2^T \mathbf{Z}_i + \epsilon_i, \end{cases} \tag{1}$$

where $\text{logit}(p) = \log(p/(1-p))$. $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ describe the genetic association of $\mathbf{G}$ and the covariate effect of $\mathbf{Z}$ on the probability of having the disease, respectively. $\boldsymbol{\beta}_3$ describes the effect of the secondary phenotype on the disease status. $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ describe the genetic association of $\mathbf{G}$ and the covariate effect of $\mathbf{Z}$ on the secondary phenotype W, and ϵ_i is the mean zero random error with variance σ^2. $I(A)$ is the indicator function of a set A, and the superscript ^{T} denotes the transpose of a vector or a matrix. For each genetic marker, the covariate G is a variable that depends on the assumed genetic model. Depending on the number of copies of the mutant allele at a particular SNP, we have $G = (G_0, G_1, G_2) = (0, 1, 1)$ for a dominant genetic model, $G = (G_0, G_1, G_2) = (0, 0, 1)$ for a recessive genetic model and $G = (G_0, G_1, G_2) = (0, 1, 2)$ for an additive genetic model [19]. Note that a conditional approach can be directly connected to Gibbs sampling in view of the relationship between the conditional distribution and joint distribution. However, Equation (1) is valid only when evaluated in the population, as the test of γ_1 in the second equation can be invalid in case-only data, because of the missingness of the secondary phenotype in the control group. In addition, assuming that the gene and other covariates affect the disease status via a secondary phenotype, *i.e.*, $P(\mathbf{D}|\mathbf{W}, \mathbf{G}, \mathbf{Z}) = P(\mathbf{D}|\mathbf{W})$, the logit model can be simplified as follows:

$$\text{logit}(P(D_i = 1|W_i)) = \beta_0 + \boldsymbol{\beta}_1^T W_i. \tag{2}$$

When W is observed for every individual, the regression coefficients in Model (1) can be estimated and tested using the maximum likelihood method. However, W values for the entire control set are only known in that they are lower in control subjects than in case subjects. If the case-control samples are obtained by 1-1 matching, then, as an example, half of the W values for the entire sample would be missing. Hence, we consider a hierarchical approach and propose Bayesian Gibbs sampling to incorporate the available information from the high correlation with disease, but only partially-observed secondary phenotype values for the case group.

An important assumption of our modeling approach is that the disease is categorized by the binary trait D, and the disease severity is measured by the secondary phenotype W for the case group. The parameter β_1 describes the associations of $\mathbf{G}_i$ with the disease categories, whereas γ_1 describes the genetic associations of the genetic component $\mathbf{G}_i$ with disease severity. Thus, to evaluate the associations of $\mathbf{G}_i$ with disease categories and severity, we test the following null and alternative hypotheses:

$$\begin{aligned} H_0^{\beta} &: \beta_1 = 0 \ \text{ versus } \ H_1^{\beta} : \beta_1 \neq 0 \\ H_0^{\gamma} &: \gamma_1 = \mathbf{0} \ \text{ versus } \ H_1^{\gamma} : \gamma_1 \neq \mathbf{0} \ , \end{aligned} \tag{3}$$

where $\mathbf{0}$ denotes the column vector of appropriate length. Testing H_0^{β} *vs.* H_1^{β} tests the effect of the secondary phenotype on the disease status, and testing H_0^{γ} *vs.* H_1^{γ} tests the existence of a genetic effect on the secondary phenotype.

5. Bayesian Testing

Based on the hierarchical approach, the corresponding joint likelihood function can be written as $L(\beta_0, \beta_1, \gamma_0, \gamma_1, \gamma_2, \sigma^2) = \prod_{i=1}^{n} pr(D_i|W_i, \beta_0, \beta_1) pr(W_i|\gamma_0, \gamma_1, \gamma_2, \sigma^2)$. For r case subjects, we have that $pr(D_i|W_i, \beta_0, \beta_1) = pr(D_i = 1|W_i, \beta_0, \beta_1)$, and for the remaining $n - r$ control subjects, we have that $pr(D_i|W_i, \beta_0, \beta_1) = pr(D_i = 0|W_i^m, \beta_0, \beta_1)$, where W_i^m is the missing phenotype for i-th subjects in the control group. Using similar representation for $pr(W_i|\gamma_0, \gamma_1, \gamma_2, \sigma^2)$, the joint likelihood can be written as:

$$\begin{aligned} L(\beta_0, \beta_1, \gamma_0, \gamma_1, \gamma_2, \sigma^2) &= \prod_{i=1}^{n} pr(D_i|W_i, \beta_0, \beta_1) pr(W_i|\gamma_0, \gamma_1, \gamma_2, \sigma^2) \\ &= \prod_{i=1}^{r} pr(D_i = 1|W_i, \beta_0, \beta_1)^{D_i} \prod_{i=r+1}^{n} pr(D_i = 0|W_i^m, \beta_0, \beta_1)^{1-D_i} \\ &\times \prod_{i=1}^{r} pr(W_i|\gamma_0, \gamma_1, \gamma_2, \sigma^2) \prod_{i=r+1}^{n} pr(W_i^m|\gamma_0, \gamma_1, \gamma_2, \sigma^2) \\ &= \prod_{i=1}^{n} \left(\frac{1}{1 + \exp(\beta_0 + \beta_1 W_i)} \right)^{D_i} \left(\frac{\exp(\beta_0 + \beta_1 W_i^m)}{1 + \exp(\beta_0 + \beta_1 W_i^m)} \right)^{1-D_i} \\ &\times \prod_{i=1}^{r} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(W_i - \gamma_0 - \gamma_1^T \mathbf{G}_i - \gamma_2^T \mathbf{Z}_i)^2}{2\sigma^2} \right) \\ &\times \prod_{i=r+1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(W_i^m - \gamma_0 - \gamma_1^T \mathbf{G}_i - \gamma_2^T \mathbf{Z}_i)^2}{2\sigma^2} \right), \end{aligned} \tag{4}$$

where r is the number of case subjects and W^m denotes the unobserved secondary phenotype, *i.e.*, a collection of secondary phenotypes for control subjects. However, under the conditional approach, the corresponding joint likelihood function cannot be expressed as an explicit form.

It has been demonstrated that hierarchical Bayesian models are well suited to the analysis of complicated augmentation problems. We consider a hierarchical model with obvious conditional independence assumptions corresponding to the representation of a genetic association model, which can be outlined as follows:

- $[\mathbf{D} \mid \mathbf{W}, \beta_0, \beta_1]$: $\text{logit}(P(D_i = 1|W_i) = \beta_0 + \beta_1 W_i$.
- $[\mathbf{W}|\mathbf{G}, \mathbf{Z}, \gamma_0, \gamma_1, \gamma_2]$: $W_i = \gamma_0 + \gamma_1^T \mathbf{G}_i + \gamma_2^T \mathbf{Z}_i + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.
- $[\beta_0, \beta_1]$: β_0 and β_1 are independent, and β_0 has an improper uniform prior; we set a relatively strong prior for β_1 to have anti-CCP values within the normal range for the control group.
- $[\gamma_0, \gamma_1, \gamma_2, \sigma^2]$: γ_0, γ_1 and γ_2 are independent and have improper, uniform priors, and σ^2 has a diffuse inverse gamma distribution.

In Bayesian model selection or a testing problem, use of the Bayes factor under proper priors or informative priors has been very successful. A summary number, which may be singled out for its clarity, is the Bayes factor [20,21]. A Bayes factor is a Bayesian alternative to frequentist hypothesis testing that is most often used for the comparison of multiple models. One reason for computing the Bayes factor is that it is based on weighing the alternative models by the posterior evidence in favor of each of them. Such evidence is not measured by the p-value of a classical hypothesis test. A small p-value provides some evidence against a null hypothesis [20,22], but a large p-value does not provide evidence in favor of the null. A second reason for computing the Bayes factor is that it can be used when comparing non-nested models. This makes the Bayes factor particularly suitable for use in constrained mixture models, where alternative models are non-nested [23]. In order to test the genetic association, we computed the Bayes factor from the Markov chain-Monte Carlo simulation of the posterior distribution.

Suppose that hypotheses $H_0^\beta : \beta_1 = 0$ *vs.* $H_1^\beta : \beta_1 \neq 0$ and $H_0^\gamma : \gamma_1 = \mathbf{0}$ vs. $H_1^\gamma : \gamma_1 \neq \mathbf{0}$ are under consideration for each genetic covariate (SNP). We let $\boldsymbol{\theta}_i = (\mathbf{W}_m, \beta_0, \beta_1, \gamma_0, \gamma_1, \gamma_2, \sigma^2)$ denote the unknown parameter under the hypothesis H_i, $i = 0, 1$. With observed data $\mathbf{x} = (\mathbf{D}, \mathbf{W} \setminus \mathbf{W}_m)$, we have probability density function $f_i(\mathbf{x}|\boldsymbol{\theta}_i)$, obtained from (4), under hypothesis H_i for $i = 0, 1$. Let $\pi_i(\boldsymbol{\theta}_i)$ be the prior distribution of hypothesis H_i, and let p_i be the prior probability of hypothesis H_i. Then, the Bayes factor of hypothesis H_1 to hypothesis H_0 is defined by:

$$B_{10} = \frac{f_1(\mathbf{x}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1}{f_0(\mathbf{x}|\boldsymbol{\theta}_0)\pi_0(\boldsymbol{\theta}_0)d\boldsymbol{\theta}_0} = \frac{m_1(\mathbf{x})}{m_0(\mathbf{x})}. \tag{5}$$

In addition, the posterior distribution that the hypothesis H_0 is true is:

$$P(H_0|\mathbf{x}) = \left(\sum_{j=0}^{1} \frac{p_j}{p_0} B_{j0}\right)^{-1} = \frac{p_0 + p_1 B_{10}}{p_0}. \tag{6}$$

6. RA Data Revisited

We next applied the hierarchical Bayesian approach to the RA data introduced in Section 2. As a preparatory step, we take the log transformation on anti-CCP (W) to reduce the skewness of the quantitative phenotype values. The posterior distribution of each parameter in Model (5) is obtained by Markov chain-Monte Carlo simulation, and Bayes factors for testing the hypotheses in (3) are calculated. The results are based on 10,000 samples after 10,000 burn-in iterations for each SNP. Table 2 summarizes the posterior mean and standard deviation of model parameters for each individual SNP presented in Table 1. In Table 2, SNPs are reordered by the magnitude of the Bayes factor for testing the null hypothesis $H_0 : \gamma_1 = 0$. The Bayes factors for testing the null hypothesis $H_0 : \gamma_1 = 0$ for each SNP are quite different from each other, whereas the Bayes factors for testing the null hypothesis $H_0 : \beta_1 = 0$ are quite similar; therefore, we ranked the SNPs by the Bayes factor for testing $H_0 : \gamma_1 = 0$. Bayes factor means the ratio of marginal likelihood under the null hypothesis and marginal likelihood under the alternative hypothesis. Kass *et al*. [21] propose the interpretation that a Bayes factor greater than three means "positive" evidence; a Bayes factor greater than 20 means "strong" evidence; and Bayes a factor greater than 150 means "very strong" evidence.

Very large Bayes factors (BFs) are not unexpected, because we have selected the top ranked 13 SNPs in the preliminary analysis. As explained in the Introduction, our setup is that secondary

phenotypes are completely missing in the control group due to different sources of data. In such a case, we cannot apply regression analysis, and we proposed the Bayesian approach as a reasonable option to model the data. We present the analysis of the RA data primarily as an example illustrating the potential of the Bayesian approach based on joint likelihood when secondary phenotypes are completely missing in the control group, except the knowledge that secondary phenotype values are lower for the control group than in the case group.

Table 2. Estimated coefficients and standard errors of model parameters for the 13 most significant single nucleotide polymorphisms (SNPs) from the logistic regression under the additive genetic model. log(BF) is log of the Bayes factor for testing $H_0 : \gamma_1 = 0$.

Rank	SNP	β_0	β_1	γ_0	γ_1	γ_2	log(BF)	γ_1^*
1	rs9442372	−5.183	0.112	3.862	−1.183	0.612	>500	−0.170
		(0.299)	(0.008)	(0.082)	(0.083)	(0.105)		(0.132)
2	rs2986742	−5.602	0.150	1.912	−1.747	2.013	>500	−0.135
		(0.394)	(0.012)	(0.236)	(0.101)	(0.235)		(0.130)
3	rs12027585	−6.574	0.174	4.061	−0.884	−0.841	477.02	−0.254
		(0.420)	(0.013)	(0.067)	(0.076)	(0.149)		(0.124)
4	rs1046269	−5.393	0.133	4.112	−1.864	−0.443	424.66	−0.143
		(0.351)	(0.010)	(0.083)	(0.095)	(0.197)		(0.120)
5	rs6671416	−5.441	0.142	4.173	−1.872	−1.044	400.90	−0.241
		(0.363)	(0.010)	(0.084)	(0.097)	(0.180)		(0.108)
6	rs356116	−7.482	0.235	3.781	−1.291	−1.636	354.24	−0.123
		(0.659)	(0.022)	(0.085)	(0.098)	(0.185)		(0.111)
7	rs6598886	−4.153	0.083	2.544	−2.048	1.756	334.60	−0.201
		(0.220)	(0.005)	(0.312)	(0.093)	(0.312)		(0.092)
8	rs16861613	−4.562	0.102	4.037	−1.638	−2.283	315.28	−0.173
		(0.259)	(0.006)	(0.081)	(0.095)	(0.248)		(0.110)
9	rs11578154	−5.353	0.135	4.083	−1.393	−1.557	299.98	−0.153
		(0.339)	(0.009)	(0.074)	(0.086)	(0.227)		(0.109)
10	rs6427128	−5.792	0.123	4.222	−0.638	−0.472	191.15	−0.207
		(0.327)	(0.008)	(0.057)	(0.065)	(0.127)		(0.109)
11	rs2062629	−8.913	0.265	3.981	−0.469	−1.013	189.78	−0.125
		(0.623)	(0.021)	(0.061)	(0.069)	(0.131)		(0.103)
12	rs7524233	−4.861	0.104	4.213	−0.888	0.244	160.42	−0.218
		(0.251)	(0.006)	(0.059)	(0.067)	(0.127)		(0.104)
13	rs2476601	−8.080	0.234	3.932	−0.809	0.786	17.12	−0.166
		(0.563)	(0.018)	(0.062)	(0.071)	(0.137)		(0.102)

γ_1^*, Estimated coefficients and standard errors for the second model in (1) using the case data only.

For a comparison with the analysis using only available data, we fit the second model of Model (1) using the case data only. We cannot fit the first logistic regression model because W_i's, *i.e.*, secondary phenotypes, are completely missing for the control group. When we compare the estimated coefficients and standard errors for γ_1 and γ_1^* in Table 2, we observe that the estimated coefficients have the same sign, but significances seem to be higher for our approach.

Furthermore, our method allows for the generation of missing secondary phenotype outcomes in the control group. The reference ranges for blood tests of anti-citrullinated protein antibodies is less than 20 for RA-negative individuals [24]. We used a strong prior for β_1 to incorporate the available knowledge on the missing phenotype values for the control group. Figure 1 shows the box plots of W_i's for the case group and the control group for the top two-ranked SNPs: "rs9442372" and "rs2986742". The median of observed anti-CCP values for the case group was 135, and the medians of missing anti-CCP values for the control group were 14.37 and 2.853 for "rs9442372" and "rs2986742", respectively.

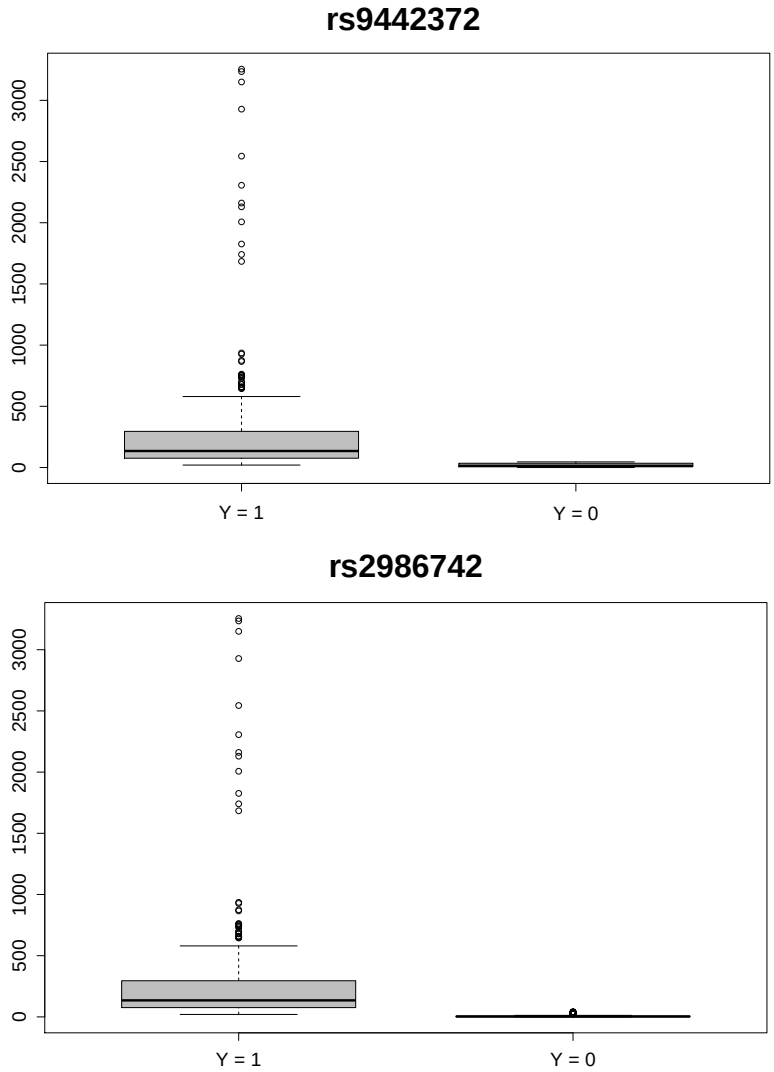

Figure 1. Box plots of anti-cyclic citrullinated peptide (CCP) values of the case ($Y = 1$) and control ($Y = 0$) groups for the two top-ranked single nucleotide polymorphisms: "rs9442372" (**upper panel**) and "rs2986742" (**lower panel**).

In Table 3, we provide the highest posterior credible sets for β_1 and γ_1, which represent the effect of the secondary phenotype on the disease status and the genetic effect on the secondary phenotype, respectively. In Bayesian inference, "credible set" terms are generally used instead of "confidence interval". Actually, a credible set can be obtained by posterior distribution. The highest posterior density interval is defined by (L, U) minimizing the length of interval satisfying $P(L < \theta < U|data) = 1 - \alpha$.

Posterior density distributions of β_1 and γ_1 for the top two ranked SNPs, "rs9442372" and "rs2986742", are plotted in Figures 2 and 3. For example, for the top ranked SNP "rs9442372", the posterior means for β_1 and γ_1 are 0.113 and -1.193, respectively. The positive posterior mean for β_1 implies that knowing the anti-CCP values seems to increase the chance of RA. The magnitude of the posterior mean for γ_1 implies that the SNP exerts quite a large genetic effect on the anti-CCP value.

Table 3. Highest posterior density credible sets for the parameters of interest (β_1, γ_1).

Rank	SNP	β_1	γ_1
1	rs9442372	$(0.110, 0.115)$	$(-1.210, -1.175)$
2	rs2986742	$(0.146, 0.155)$	$(-1.802, -1.692)$
3	rs12027585	$(0.171, 0.182)$	$(-0.906, -0.862)$
4	rs1046269	$(0.131, 0.138)$	$(-1.888, -1.840)$
5	rs6671416	$(0.135, 0.142)$	$(-1.900, -1.843)$
6	rs356116	$(0.219, 0.240)$	$(-1.345, -1.236)$
7	rs6598886	$(0.083, 0.085)$	$(-2.074, -2.021)$
8	rs16861613	$(0.100, 0.103)$	$(-1.686 - 1.590)$
9	rs11578154	$(0.130, 0.135)$	$(-1.426, -1.360)$
10	rs6427128	$(0.120, 0.125)$	$(-0.650, -0.625)$
11	rs2062629	$(0.251, 0.274)$	$(-0.499, -0.438)$
12	rs7524233	$(0.094, 0.097)$	$(-0.901, -0.874)$
13	rs2476601	$(0.225, 0.242)$	$(-0.825, -0.793)$

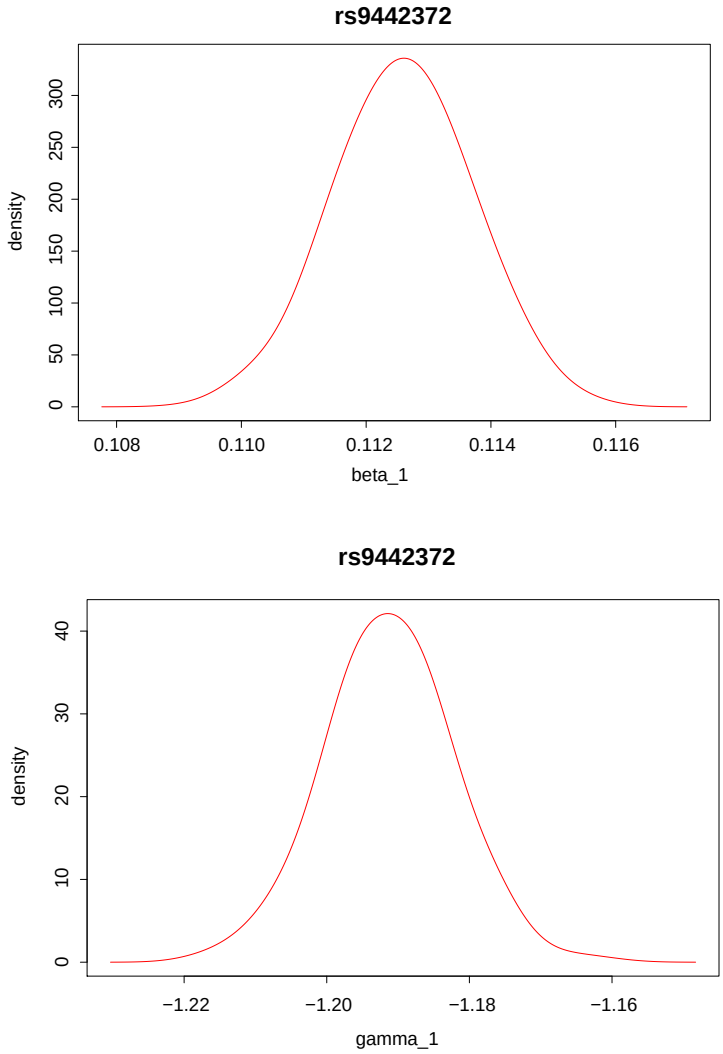

Figure 2. Posterior distribution of β_1 (**upper panel**) and γ_1 (**lower panel**) for single nucleotide polymorphism "rs9442372", which has the largest Bayes factor.

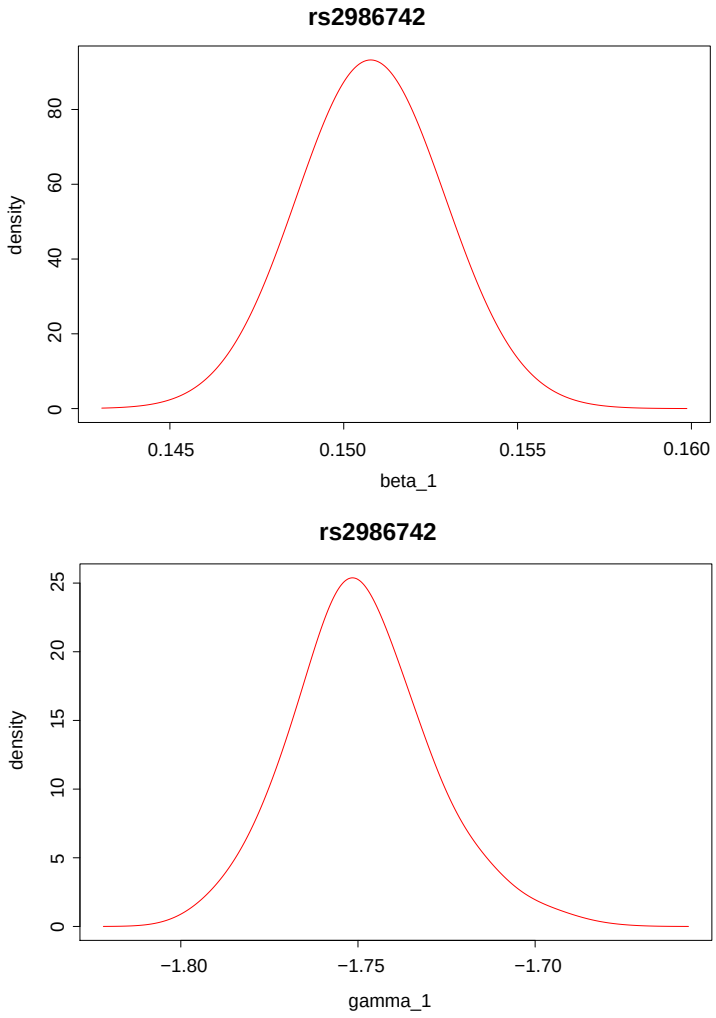

Figure 3. Posterior distribution of β_1 (**upper panel**) and γ_1 (**lower panel**) for single nucleotide polymorphism "rs2986742", which has the second largest Bayes factor.

7. Discussion

When information on the secondary phenotype is available only for the case group due to cost and different data sources in genome-wide association studies, it is not feasible to fit a linear regression incorporating stratum-specific missing data. On the other hand, fitting a linear regression model ignoring supplementary phenotype data may provide limited knowledge regarding genetic association. In this paper, we have considered Bayesian genetic association testing when only partial secondary continuous phenotype trait values are available and investigated the genetic effects on disease status through the intermediate secondary phenotype. We illustrate the use of Bayes factors in GWAS for selected SNPs whose p-values from logistic regression are highly significant. Using a Bayesian approach for inference, available but incomplete information is reflected in the model through the prior established on the parameter. For example, we set a strong prior on β_1 to make use of the knowledge that the partially-observed anti-CCP values are lower for the control group than the case group. In addition, we were able to incorporate the partially-observed secondary outcomes related to the disease in the model and generated the missing outcomes in the control group.

Our setup is that secondary phenotypes are completely missing in the control group due to different sources of data. In such a case, we cannot apply regression analysis, and we proposed the Bayesian approach as a reasonable option to model the data. We present the analysis of the RA data primarily as an example illustrating the potential of the Bayesian approach based on joint likelihood

when secondary phenotypes are completely missing in the control group, except the knowledge that secondary phenotype values are lower for the control group than in the case group. Our approach can be applicable to the whole genome-wide association study. The running time per SNP was about 480 s per 20,000 iterations using R Version 3.1.3 on a personal computer (Intel Core I-7 at 3.40 GHZ).

With the liability threshold model, Falconer [25] supposed that there is a hypothetical and continuous attribute, which he referred to as the individual's "liability" to the disease of interest, and all individuals above a certain threshold are affected with the disease. This notion of liability is similar to a latent variable in statistics, and his assumption is limited to normally distributed liability with equal variance for the case and the control groups. However, our motivation was to model "mediated pleiotropy", where a causal gene affects one phenotype (anti-CCP) that lies on the causal path to another phenotype (RA status), and thus, an association occurs between the observed gene and both phenotypes. There are multivariate analyses jointly analyzing more than one phenotype in a unified framework and test for the association of multiple phenotypes with a genetic variant, for example multivariate analysis of variance (MANOVA). However, MANOVA or other ordinary regression approaches cannot handle the case when the first phenotype is not available for the whole control group. Our approach handles such a case, and the Bayesian framework allows us to test the association between a genetic variant and the phenotype of interest making use of partial information on the missing phenotypes in the control group. We present the results based on real data analysis to illustrate the potential of the Bayesian approach proposed in this paper. The signal presented in our table should be interpreted with caution, and it would be nice to validate it using independent data sample.

In genome-wide association studies, there is a variety of efforts to recruit more study subjects, because we need to test for a huge number of genetic variants with a rather small number of subjects compared to the number of SNPs. This may include collecting data at both population and family levels and combining data from different resources. Ascertainment bias happens when there is more intensive screening for the outcome among the affected than among the unaffected. Then, the case-control ratio in the available sample does not necessarily represent the prevalence of the disease at the population level due to the way the data are collected. There are several attempts to explore or explicitly incorporate ascertainment bias, for example conducting sensitivity analysis or conditional likelihood approach (Lachance and Tishkoff [26] and Haghighi and Hodge [27]). We leave the issue of incorporating ascertainment bias in Bayesian framework as future work.

The use of a fully-Bayesian approach for hypothesis testing eliminates the difficulties of constructing estimators, because no estimation is required; instead, hypothesis tests are obtained directly from the likelihood, with nuisance parameters integrated out. If parameter estimates are required, they can be obtained from the posterior distribution as posterior means and credible intervals, as demonstrated in this article.

The Bayes factor is a summary measure that provides an alternative to the p-value for the ranking of associations or the flagging of associations as significant. Andrews [28] showed the relationship between the Bayes factors and Wald, likelihood ratio and score statistics under quite general priors. Instead of reporting p-values, whose interpretation depends on the sample size, we recommend reporting Bayes factors with p-values. For a given test statistic, p-values only measure significance, whereas Bayes factors integrate both the significance and the sample size to detect an association. One benefit of reporting Bayes factors is that they are more appropriate measures than p-values when comparing results across GWAS with different sample sizes, which is a frequent situation in GWAS for common diseases.

The case subjects that we made available for our analysis comprise independent individuals who have met the American College of Rheumatology criteria for rheumatoid arthritis. These cases comprise a single member of 445 sib-pairs that were studied as a part of the North American Rheumatoid Arthritic Consortium and an additional 423 cases who were not selected for family

history. The cases were recruited from across the United States. Cases are predominantly of Northern European origin. The control subjects, derived from the New York Cancer Projects, were enrolled in the New York metropolitan area. These controls are enriched for individuals of Southern European or Ashkenazi Jewish ancestry compared to cases [12]. The problem of non-random ascertainment has been usually dealt with by formulating a conditional likelihood, which leads to the removal of some individuals from the study in order to avoid an outcome-dependent ascertainment bias. With our data, we did not think that our results are affected by case-control ascertainment, because the unrelated control group was obtained quite independently from the case subjects. Population stratification due to different ancestries within Europe might exist in our case-control data, and incorporating such stratification in the analysis is beyond the scope of this paper.

Acknowledgments: This research was supported by the year 2015 Yeungnam University Research Grant. This work is based on the GAW16 data gathered with the support of grants from the National Institutes of Health (NO1-AR-2-2263 and RO1-AR-44422, Peter K. Gregersen, PI) and the National Arthritis Foundation.

Author Contributions: Minjung Kwak conceived and designed the research; Yongku Kim and Minjung Kwak analyzed the data and interpreted the results; Yongku Kim and Minjung Kwak wrote and revised the manuscript. Both authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Draper, N.R.; Smith, H. *Applied Regression Analysis*, 3rd ed.; Wiley: New York, NY, USA, 1998.
2. Chatterjee, N.; Carroll, R.J. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **2005**, *92*, 399–418.
3. Agresti, A. *Categorical Data Analysis*; Wiley: New York, NY, USA, 1990.
4. Zheng, G.; Freidlin, B.; Gastwirth, J.L. Robust genomic control for association studies. *Am. J. Hum. Genet.* **2006**, *78*, 350–356.
5. Chatterjee, N.; Kalaylioglu, Z.; Moslehi, R.; Peters, U.; Wacholder, S. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am. J. Hum. Genet.* **2006**, *79*, 1002–1016.
6. Maity, A.; Carroll, R.J.; Mammen, E.; Chatterjee, N. Testing in semiparametric models and interaction, with applications to gene-environment interactions. *J. R. Statist. Soc. B* **2009**, *71*, 75–96,
7. The Welcome Trust Case Control Consortium (WTCCC). Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **2007**, *447*, 661–678.
8. Wu, C.; Zheng, G.; Kwak, M. A joint regression analysis for genetic association studies with outcome stratified samples. *Biometrics* **2013**, *69*, 417–426.
9. He, J.; Li, H.; Edmondson, A.C.; Rader, D.J.; Li, M. A Gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics* **2012**, *13*, 497–508.
10. Stephens, M.; Balding, D.J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **2009**, *10*, 681–690.
11. Xu, J.; Yuan, A.; Zheng, G. Bayes factor based on the trend test incorporating Hardy-Weinberg disequilibrium: More powerful to detext genetic association. *Ann. Hum. Genet.* **2012**, *76*, 301–311.
12. Amos, C.I.; Chen, W.V.; Seldin, M.F.; Remmers, E.F.; Taylor, K.E.; Criswell, L.A.; Lee, A.T.; Plenge, R.M.; Kastner, D.L.; Gregersen, P.K. Data for genetic analysis workshop 16 Problem 1, association analysis of rheumatoid arthritis data. *BMC Proc.* **2009**, *3*, doi:10.1186/1753-6561-3-S7-S2.
13. Worthington, J.; Barton, A.; John, S.L. *The Epidemiology of Rheumatoid Arthritis and the Use of Linkage and Association Studies to Identify Disease Genes*; Springer-Birkhäuser: Basel, Switzerland, 2005.
14. Huizinga, T.W.; Amos, C.I.; van der Helm-van Mil, A.H.; Chen, W.; van Gaalen, F.A.; Jawaheer, D.; Schreuder, G.M.; Wener, M.; Breedveld, F.C.; Ahmad, N.; *et al*. Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins. *Arthritis Rheum.* **2005**, *52*, 3433–3438.
15. Kroot, E.J.; de Jong, B.A.; van Leeuwen, M.A.; Swinkels, H.; van den Hoogen, F.H.; van't Hof, M.; van de Putte, L.B.; van Rijswijk, M.H.; van Venrooij, W.J.; van Riel, P.L. The prognostic value of anti-cyclic

citrullinated peptide antibody in patients with recent-onset rheumatoid arthritis. *Arthritis Rheum.* **2000**, *43*, 1831–1835.
16. Chen, L.; Zhong, M.; Chen, W.V.; Amos, C.I.; Fan, R. A genome-wide association scan for rheumatoid arthritis data by Hotelling's T^2 tests. *BMC Proc.* **2009**, *3*, doi:10.1186/1753-6561-3-S7-S6.
17. Suzuki, A.; Yamada, R.; Chang, X.; Tokuhiro, S.; Sawada, T.; Suzuki, M.; Nagasaki, M.; Nakayama-Hamada, M.; Kawaida, R.; Ono, M.; *et al*. Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.* **2003**, *34*, 395–402.
18. Begovich, A.B.; Carlton, V.E.; Honigberg, L.A.; Schrodi, S.J.; Chokkalingam, A.P.; Alexander, H.C.; Ardlie, K.G.; Huang, Q.; Smith, A.M.; Spoerke, J.M.; *et al*. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.* **2004**, *75*, 330–337.
19. Sasieni, P.D. From genotypes to genes: Doubling the sample size. *Biometrics* **1997**, *53*, 1253–1261.
20. Berger, J.O.; Sellke, T. Testing a point null hypothesis: The irreconcilability of *p* values and evidence. *J. Am. Stat. Assoc.* **1987**, *82*, 112–122.
21. Kass, R.E.; Raftery, A.E. Bayes factors. *J. Am. Stat. Assoc.* **1995**, *90*, 773–795.
22. Casella, G.; Berger, R.L. Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion). *J. Am. Stat. Assoc.* **1987**, *82*, 106–111.
23. Clogg, C.C.; Goodman, L.A. Latent structure analysis of a set of multidimensional contingency tables. *J. Am. Stat. Assoc.* **1984**, *79*, 762–771.
24. Coenen, D.; Verschueren, P.; Westhovens, R.; Bossuyt, X. Technical and diagnostic performance of 6 assays for the measurement of citrullinated protein/peptide antibodies in the diagnosis of rheumatoid arthritis. *Clin. Chem.* **2007**, *53*, 498–504.
25. Falconer, D.S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **1965**, *29*, 51–76.
26. Lachance, J.; Tishkoff, S.A. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *Bioessays* **2013**, *35*, 780–786.
27. Haghighi, F.; Hodge, S.E. Likelihood formulation of parent-of-origin effects on segregation analysis, including ascertainment. *Am. J. Hum. Genet.* **2002**, *70*, 142–156.
28. Andrews, D.W.K. The large sample correspondence between classical hypothesis tests and Bayesian posterior odds tests. *Econometrica* **1994**, *62*, 1207–1232.