Supplementary notes for "Exploring predictive biomarkers from clinical genome-wide association studies via multidimensional hierarchical mixture models"

Takahiro Otani, Hisashi Noma, Shonosuke Sugasawa, Aya Kuchiba, Atsushi Goto, Taiki Yamaji, Yuta Kochi, Motoki Iwasaki, Shigeyuki Matsui, and Tatsuhiko Tsunoda

Contents

A. Descriptions of randomized clinical trials

The two datasets used in this analysis were deposited in the dbGaP database and derived from the VISP trial (study accession number: phs000343.v3.p1) and the SUCCESS-A trial (study accession number: phs000547.v1.p1).

The VISP trial was a multi-center, double-blind, randomized, controlled clinical trial that enrolled patients aged 35 or older with homocysteine levels above the 25th percentile at screening and a non-disabling cerebral infarction within 120 days of randomization. The trial was designed to determine if daily intake of a multivitamin tablet containing high-dose folic acid, vitamin B6, and vitamin B12 reduced recurrent cerebral infarction (primary endpoint) as well as nonfatal myocardial infarction or mortality (secondary endpoints). Subjects were randomly assigned to receive daily doses of the high-dose formulation, containing 25 mg pyridoxine (B6), 0.4 mg cobalamin (B12), and 2.5 mg folic acid (treatment group); or the low-dose formulation, containing 200 mcg pyridoxine, 6 mcg cobalamin, and 20 mcg folic acid (control group). Within the trial, 2,164

participants from 46 clinic sites provided DNA and agreed for it to be shared for use in a genetic subset study of VISP, and 1051295 SNPs were genotyped using the Illumina HumanOmni1-Quad_v1-0_B BeadChip. The resulting dataset was analyzed using a previous approach[1] to investigate SNPs associated with blood homocysteine levels, which are strongly associated with the risk of cardiovascular disease.

The SUCCESS-A trial was a randomized phase III study of treatment response of early primary breast cancer to adjuvant therapy after surgical resection. The trial was designed to determine if adjuvant chemotherapy with gemcitabine, an antimetabolite frequently used in the treatment of pancreatic cancer and other diseases[2], improved progression-free survival, overall survival, and toxicity. Subjects were randomly assigned to chemotherapy with gemcitabine (treatment group) or without gemcitabine (control group). The treatment group received chemotherapy with gemcitabine (three cycles of 5-fluorouracil 500 mg/m² i.v. body surface area, epirubicin 100 mg/m² i.v., and cyclophosphamide 500 mg/m² i.v. (FEC100), each administered on day 1 and repeated on day 22, subsequently followed by three cycles of docetaxel 75 mg/m² body surface area i.v., and gemcitabine 1000 mg/m² i.v. (30 min infusion), administered on day 1, followed by gemcitabine 1000 mg/m² i.v. (30 min infusion) on day 8, repeated on day 22), and the control group received chemotherapy without gemcitabine (three cycles of FEC100, each administered on day 1 and repeated on day 22, subsequently followed by three cycles of docetaxel 100 mg/m² body surface area i.v., administered on day 1 and repeated on day 22). A total of 3322 participants from 250 clinic sites across Germany provided DNA, and 693543 SNPs were successfully genotyped using the Illumina HumanOmniExpress-FFPE BeadChip. In this study, we used this dataset to investigate SNPs associated with progression-free survival in breast cancer patients.

B. Quality control

For quality control in the VISP trial dataset we excluded the following: (i) subjects with no homocysteine level data, (ii) non-whites, (iii) an individual with an outlying homocysteine level (using the procedure of Wakefield et al.[1]), (iv) subjects with no genotype data, (v) SNPs with genotype call rates of <95%, (vi) subjects with subject call rates of <98%, (vii) SNPs with genotype call rates calculated from remained subjects of <98%, (viii) SNPs that deviated from Hardy-Weinberg equilibrium ($p < 10^{-6}$), (ix)

SNPs with a minor allele frequency (MAF) of <1%, and (x) sex chromosome SNPs. A total of 1533 subjects (760 assigned to the treatment group and 773 assigned to the control group) with 774670 SNPs passed this process.

The SUCCESS-A trial dataset passed a quality control filter defined by the data provider. We used "MAF-event"-filtered genotype data included in the dbGaP dataset that excluded SNPs with $2 \times n \times \text{MAF} \times (1 - \text{MAF}) \leq 75$, where n is the number of disease progression events, and excluded sex chromosome SNPs for the main analysis using the hierarchical mixture model. As in a preliminary association analysis on progression-free survival conducted by the data provider, we also excluded (i) HapMap control subjects, (ii) related subjects, (iii) subjects with subject call rates of <95%, and (iv) principal components analysis (PCA)-defined Asian subjects. A total of 3289 subjects (1621 subjects assigned to the treatment group and 1668 subjects assigned to the control group) with 424121 SNPs passed this process.

C. Association analysis

For the VISP trial, the outcome was the difference in blood homocysteine levels between baseline and the first post-baseline measurements, as in the study of Wakefield et al.[1]; this outcome was used to compare the effectiveness of standard methods and the new method, although homocysteine levels were measured longitudinally in the trial. Let $Y_{i,j}$ represent the SNP-specific outcome for individual $i = 1, \ldots, N$ and SNP $j = 1, \ldots, M$. For control and treatment groups, we assumed the following linear regression model:

$$Y_{i,j} = \alpha_j + \boldsymbol{x}_i'\boldsymbol{\phi}_j + G_{i,j}\beta_j + \varepsilon_i,$$

where $\boldsymbol{x}_i$ corresponds to individual-level covariates consisting of age and gender, and $G_{i,j}$ is the number of reference alleles. α_j, $\boldsymbol{\phi}_j$, and β_j are regression coefficients and are estimated via maximum likelihood estimation. Further, ε_i is an error term with a mean of zero and a variance of σ^2.

For the SUCCESS-A trial, we conducted an association analysis using the same approach as that employed in the preliminary association tests that are included in the dbGaP dataset of the trial. There were 341 events in 3,289 unrelated patients. For control and treatment groups, we assumed a proportional hazards regression model for all patients stratified by overall estrogen receptor status (positive or negative) and type of DNA

sample (DNA from original blood sample, restored DNA from original blood sample, or new blood sample):

$$h_{i,j}(t) = h_{0,j}(t)\exp\left(\boldsymbol{x}_i'\boldsymbol{\phi}_j + G_{i,j}\beta_j\right),$$

where $h_{0,j}(t)$ corresponds to a SNP-specific baseline hazard. The covariates $\boldsymbol{x}_i$ consist of age, body mass index (BMI), tumor grading, tumor stage pN and pT according to TNM classification, and two numeric scores for the first and second components from PCA to adjust for possible population stratification.

D. Multi-subgroup gene screening via hierarchical mixture models

Multidimensional hierarchical mixture modeling

We used an empirical Bayes framework based on a multidimensional semi-parametric hierarchical mixture model proposed by Matsui et al.[3] to estimate the underlying distribution of SNPs that are associated with outcomes. An R program of the multidimensional hierarchical mixture models is available at http://normanh.skr.jp/software/crestbigdata.html. Suppose that M SNPs are simultaneously tested to determine whether either is associated with disease risk. Of these SNPs, M_0 are truly "null" and are not associated with outcomes, and $M_1 = M - M_0$ SNPs are truly "non-null" and are associated with outcomes.

We defined an effect size vector $\boldsymbol{\beta}_j = \left(\beta_j^{(0)}, \beta_j^{(1)}\right)$ for SNP j as the regression coefficient for the reference allele under an additive genetic model, where $\beta_j^{(0)}$ and $\beta_j^{(1)}$ are coefficients for control and treatment groups. As an estimate of $\boldsymbol{\beta}_j$, we considered the maximum likelihood estimate $\boldsymbol{b}_j = \left(b_j^{(0)}, b_j^{(1)}\right)$. We assumed a mixture model for the $\boldsymbol{b}_j$'s,

$$f\left(\boldsymbol{b}_j, \boldsymbol{\Sigma}_j\right) = \pi f_0\left(\boldsymbol{b}_j, \boldsymbol{\Sigma}_j\right) + (1-\pi) f_1\left(\boldsymbol{b}_j, \boldsymbol{\Sigma}_j\right),$$

where f_0 and f_1 are the density functions of $\boldsymbol{b}$ for the null and non-null SNPs, and $\boldsymbol{\Sigma}_j = \mathrm{diag}\left(V_j^{(0)}, V_j^{(1)}\right)$ is a covariance matrix consisting of empirical variances (squared standard errors) from an association analysis for a particular SNP for each group. We considered different modeling assumptions for the two components. For the null component, we assume that f_0 has the normal distribution $N\left(\mathbf{0}, \boldsymbol{\Sigma}_j\right)$.

For the non-null component of interest, a hierarchical model was assumed. To be specific, for a non-null SNP j, we assumed

$$\boldsymbol{b}_j|\boldsymbol{\beta}_j \sim N\left(\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_j\right) \text{ and } \boldsymbol{\beta}_j \sim g_1(\cdot).$$

In the first level of this model, given a SNP-specific effect size $\boldsymbol{\beta}_j$, $\boldsymbol{b}_j$ follows a normal distribution. In the second level, the SNP-specific $\boldsymbol{\beta}_j$ follows the distribution g_1. The marginal distribution f_1 is given by

$$f_1(\boldsymbol{b}, \boldsymbol{\Sigma}) = \int_{-\infty}^{\infty} g_1(\boldsymbol{\beta}) \varphi_{\boldsymbol{\beta},\boldsymbol{\Sigma}}(\boldsymbol{b}) d\boldsymbol{\beta}$$

where $\varphi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\cdot)$ is the density function of the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

We estimated the parameters π and g_1 via an EM algorithm in the same way as previous studies[3–5]. We considered a nonparametric estimate of the prior distribution g_1 in which the estimate was supported by fixed discrete mass points at a series of nonzero points (the zero point was skipped because we considered non-zero effects for non-null SNPs), and an estimate of the marginal distribution f_1 was calculated using summations rather than integrations.

Detecting disease-related SNPs

To detect SNPs that are associated with outcomes, some SNP-specific indices were defined. Let γ_j be the unknown indicator variable for null/non-null status for SNP j, such that $\gamma_j = 1$ if SNP j is non-null and $\gamma_j = 0$ otherwise. The prior probability of being non-null is $\mathrm{P}(\gamma_j = 1) = 1 - \pi$ and the posterior probability is $\Pr(\gamma_j = 1|\boldsymbol{b}_j = \boldsymbol{b}, \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}) = (1-\pi) f_1(\boldsymbol{b}, \boldsymbol{\Sigma}) / f(\boldsymbol{b}, \boldsymbol{\Sigma})$. The posterior probability was estimated as

$$\tau_j = \frac{(1-\hat{\pi})\hat{f}_1\left(\boldsymbol{b}_j, \boldsymbol{\Sigma}_j\right)}{\hat{\pi} f_0\left(\boldsymbol{b}_j, \boldsymbol{\Sigma}_j\right) + (1-\hat{\pi})\hat{f}_1\left(\boldsymbol{b}_j, \boldsymbol{\Sigma}_j\right)}$$

where $\hat{\pi}$ and $\hat{f}_1$ are the empirical estimates of π and f_1. We used an ODP statistic under the empirical Bayes framework based on the hierarchical Bayesian models derived by Noma and Matsui[6] to screen disease-related SNPs. Adapting their results to the hierarchical mixture model, the ODP statistic becomes

$$R_{\mathrm{ODP}}(\boldsymbol{b}, \boldsymbol{\Sigma}) = \frac{\hat{f}_1(\boldsymbol{b}, \boldsymbol{\Sigma})}{f_0(\boldsymbol{b}, \boldsymbol{\Sigma})}.$$

Multiple hypothesis testing involving estimation of the FDR in the Bayesian sense was conducted as follows. Let k be the number of tests called significant. We first calculated R_{ODP} for each SNP. Then, we ranked the SNPs in order of decreasing R_{ODP}, so that $j = 1, \ldots, k$ represent the tests called significant. The FDR of significant results can be estimated as

$$\widehat{\mathrm{FDR}} = \frac{1}{k}\sum_{j=1}^{k}(1-\tau_j).$$

Adjusted effect size estimates

The estimated effect sizes derived from association analysis contain two types of errors, (i) a selection error, i.e., incorrectly selecting SNPs with no association; and (ii) an overestimation error. Using the estimated underlying distribution of effect sizes, we can obtain the effect size estimates adjusted for these different types of errors[7]. The selection error is adjusted by the estimated posterior probability of association τ_j. Also, the overestimation error is adjusted by an unconditional mean for the effect size for SNP j. The posterior mean for the effect size for SNP j is

$$\mathrm{E}\left(\boldsymbol{\beta}|\gamma_j = 1, \boldsymbol{b}_j = \boldsymbol{b}, \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}\right) = \int \boldsymbol{\beta} f\left(\boldsymbol{\beta}|\gamma_j = 1, \boldsymbol{b}_j = \boldsymbol{b}, \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}\right)\mathrm{d}\boldsymbol{\beta},$$

and this can be estimated as $\widehat{\mathrm{E}}(\boldsymbol{\beta}|\gamma_j = 1, \boldsymbol{b}_j = \boldsymbol{b}, \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma})$ by plugging in the estimate of f. By combining this conditional mean, given $\gamma_j = 1$, with the posterior probability of being non-null, the unconditional mean as an index for effect size for SNP j is derived as

$$\mathrm{E}\left(\boldsymbol{\beta}|\boldsymbol{b}_j = \boldsymbol{b}, \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}\right) = \Pr\left(\gamma_j = 1|\boldsymbol{b}_j = \boldsymbol{b}\right)\mathrm{E}(\boldsymbol{\beta}|\gamma_j = 1, \boldsymbol{b}_j = \boldsymbol{b}, \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}),$$

and the adjusted effect size estimate $\boldsymbol{w}_j$ is obtained by plugging in the hyperparameter estimates as

$$\boldsymbol{w}_j = \widehat{\mathrm{E}}\left(\boldsymbol{\beta}|\boldsymbol{b}_j = \boldsymbol{b}, \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}\right) = \tau_j\widehat{\mathrm{E}}(\boldsymbol{\beta}|\gamma_j = 1, \boldsymbol{b}_j = \boldsymbol{b}, \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}).$$

E. Association analyses based on conventional interaction tests

For the VISP trial, we conducted association tests on the difference in blood homocysteine levels between the baseline and the first post-baseline measurements in the same way as

Wakefield et al.[1]. Let $Y_{i,j}$ represent the SNP-specific outcome for individual $i = 1, \ldots, N$ and SNP $j = 1, \ldots, M$. We assumed the following linear regression model for SNP j:

$$Y_{i,j} = \alpha_j + \boldsymbol{x}_i'\boldsymbol{\phi}_j + G_{i,j}\beta_j + T_i\gamma_j + T_i \times G_j\Delta_j + \varepsilon_i,$$

where $\boldsymbol{x}_i$ corresponds to individual-level covariates consisting of age and gender, $G_{i,j}$ is the number of reference alleles, T_i is the indicator for treatment ($T_i = 1$ for high-dose multivitamin tablets and $T_i = 0$ for control). α_j, $\boldsymbol{\phi}_j$, β_j, γ_j, and Δ_j are regression coefficients and are estimated via maximum likelihood estimation (MLE). Further, ε_i is an error term with a mean of zero and a variance of σ^2. The null hypothesis for detecting gene-by-treatment interactions is $H_0: \Delta_j = 0$. Hypothesis testing is based on the Z statistic $Z = \widehat{\Delta}_j/\sqrt{V_j}$, with V_j as the estimated asymptotic variance of the MLE. The observed p-value based on a Z statistic is $p = \Pr(|Z| > z_{obs}|H_0)$.

For the SUCCESS-A trial, our association analysis used the same methodology as the preliminary association tests included in the dbGaP dataset of the trial. We assumed a proportional hazards regression model for all breast cancer patients stratified by overall estrogen receptor status (positive or negative) and type of DNA sample (DNA from original blood sample, restored DNA from original blood sample, or new blood sample):

$$h_{i,j}(t) = h_{0,j}(t)\exp\left(\boldsymbol{x}_i'\boldsymbol{\phi} + G_{i,j}\beta_j + T_i\gamma_j + T_i \times G_j\Delta_j\right),$$

where $h_{0,j}(t)$ corresponds to a SNP-specific baseline hazard. The covariates $\boldsymbol{x}_i$ consist of age, body mass index (BMI), tumor grading, tumor stage pN and pT according to TNM classification, and two numeric scores for the first and second components from PCA to adjust for possible population stratification. SNPs with the smallest p-values in testing the interaction effect Δ_j will be selected as disease-related SNPs.

In addition, to illustrate the efficacy of the new testing method using hierarchical mixture models, we compared the number of detected SNPs between conventional association tests using regression models with interaction terms and the ODP[6,8] with the hierarchical mixture models under specified FDR levels (**Table S10**). To estimate FDR levels in the standard analysis, we used the qvalue procedure[9] (available at https://github.com/jdstorey/qvalue) with default tuning parameters.

F. Simulation study

We assessed the performance of the ODP through a simulation study based on the stroke

trial and the breast cancer trial described in Section A. As in the study of Matsui et al.[3], we simulated a dataset with random values of effect size estimates $\boldsymbol{b}_j$ using the hierarchical mixture model described in Section D. We set the true effect size distribution g_1 as the empirical estimates derived from the analysis of two clinical trials (**Figure 2**), and the proportion of null SNPs π as 0.9, 0.99, or 0.999. We also set the covariance matrix $\boldsymbol{\Sigma}_j$ as $\mathrm{diag}\left(V_j^{(0)}, V_j^{(1)}\right)$ or $\mathrm{diag}\left(V_j^{(0)} \cdot N_c/2500,\ V_j^{(1)} \cdot N_t/2500\right)$ where N_c and N_t are sample sizes of the control group and the treatment group in the original datasets. The latter setting corresponds to the case where the same clinical trials were conducted with 5000 subjects (2500 assigned to the treatment group and 2500 assigned to the control group).

For each simulated dataset, we performed the ODP described in Section D and those corresponding to the conventional association tests. As the conventional tests, we used the standardized test statistics for SNP j,

$$S_j = \frac{\left(1/V_j^{(0)}\right) b_j^{(0)} + \left(1/V_j^{(1)}\right) b_j^{(1)}}{\sqrt{1/V_j^{(0)} + 1/V_j^{(1)}}}, T_j = \frac{b_j^{(0)} - b_j^{(1)}}{V_j^{(0)} + V_j^{(1)}}$$

for detecting prognostic and predictive SNPs, respectively. The FDR for these tests were estimated using the qvalue procedure[9] with default tuning parameters.

Tables S11 and **S12** summarize the average number of significant SNPs and true positives at FDR=5, 10 or 20% across 200 simulations. For all scenarios, the ODP consistently detected larger numbers of significant SNPs with controlling FDR accurately, compared with the conventional methods. These results indicate the efficiency of the new screening method.

References

1 Wakefield J, Skrivankova V, Hsu F-C, Sale M, Heagerty P. Detecting signals in pharmacogenomic genome-wide association studies. *Pharmacogenomics J* 2014; **14**: 309–15.

2 Soo RA, Yong W-P, Innocenti F. Systemic therapies for pancreatic cancer - the role of pharmacogenetics. *Curr Drug Targets* 2012; **13**: 811–828.

3 Matsui S, Noma H, Qu P *et al.* Multi-subgroup gene screening using semi-parametric hierarchical mixture models and the optimal discovery procedure: application to a randomized clinical trial in multiple myeloma. *Biometrics* 2017. doi:10.1111/biom.12716.

4 Matsui S, Noma H. Estimating effect sizes of differentially expressed genes for power and sample-size assessments in microarray experiments. *Biometrics* 2011; **67**: 1225–1235.

5 Nishino J, Kochi Y, Shigemizu D *et al.* Empirical Bayes estimation of semi-parametric hierarchical mixture models for unbiased characterization of polygenic disease architectures. *bioRxiv, http://biorxiv.org/lookup/doi/101101/080945* 2016. doi:10.1101/080945.

6 Noma H, Matsui S. The optimal discovery procedure in multiple significance testing: an empirical Bayes approach. *Stat Med* 2012; **31**: 165–176.

7 Matsui S, Noma H. Estimation and selection in high-dimensional genomic studies for developing molecular diagnostics. *Biostatistics* 2011; **12**: 223–233.

8 Storey JD, Dai JY, Leek JT. The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics* 2007; **8**: 414–432.

9 Storey JD. A direct approach to false discovery rates. *J R Stat Soc Ser B* 2002; **64**: 479–498.