Exploring predictive biomarkers from clinical genome-wide association studies via multidimensional hierarchical mixture models

Takahiro Otani[1], Hisashi Noma[2], Shonosuke Sugasawa[3], Aya Kuchiba[4], Atsushi Goto[5], Taiki Yamaji[5], Yuta Kochi[6], Motoki Iwasaki[5], Shigeyuki Matsui[1], and Tatsuhiko Tsunoda[7,8]

[1] Department of Biostatistics, Nagoya University Graduate School of Medicine, Nagoya, Aichi, Japan

[2] Department of Data Science, The Institute of Statistical Mathematics, Tachikawa, Tokyo, Japan

[3] Center for Spatial Information Science, The University of Tokyo, Kashiwa, Chiba, Japan

[4] Division of Biostatistical Research, Center for Public Health Sciences, National Cancer Center, Chuo-ku, Tokyo, Japan

[5] Division of Epidemiology, Center for Public Health Sciences, National Cancer Center, Chuo-ku, Tokyo, Japan

[6] Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan

[7] Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Bunkyo-ku, Tokyo, Japan

[8] Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan

Correspondence: Takahiro Otani, Ph.D.

Department of Biostatistics

Nagoya University Graduate School of Medicine

65 Tsurumai-cho, Showa-ku, Nagoya, Aichi 466-8550, Japan

Phone +81-52-744-2487

e-mail: otani@med.nagoya-u.ac.jp

Running title: Exploring biomarkers from clinical GWASs

Conflict of interest: The authors declare no conflict of interest.

Abstract

Although the detection of predictive biomarkers is of particular importance for the development of accurate molecular diagnostics, conventional statistical analyses based on gene-by-treatment interaction tests lack sufficient statistical power for this purpose, especially in large-scale clinical genome-wide studies that require an adjustment for multiplicity of a huge number of tests. Here we demonstrate an alternative efficient multi-subgroup screening method using multi-dimensional hierarchical mixture models developed to overcome this issue, with application to stroke and breast cancer randomized clinical trials with genomic data. We show that estimated effect size distributions of single nucleotide polymorphisms (SNPs) associated with outcomes, which could provide clues for exploring predictive biomarkers, optimizing individualized treatments, and understanding biological mechanisms of diseases. Furthermore, using this method we detected three new SNPs that are associated with blood homocysteine levels, which are strongly associated with the risk of stroke. We also detected six new SNPs that are associated with progression-free survival in breast cancer patients.

Keywords: predictive biomarker; randomized clinical trial; genome-wide association study; multidimensional hierarchical mixture model; optimal discovery procedure

Introduction

The development of accurate molecular diagnostics for choosing the best treatment to maximize benefits or minimize risks in a particular individual is a crucial issue for the realization of precision medicine. This will require predicting therapeutic responses, and to this end, it is of particular importance to efficiently explore predictive biomarkers that successfully classify patients so that treatment effects differ between subgroups. One of the most promising study designs for this purpose is the genome-wide association study (GWAS), as this approach allows for the investigation of medical traits such as drug metabolism, efficacy, and toxicity[1–3]. In these studies, detection of gene-by-treatment interactions is one of the crucial issues for developing predictive biomarkers[4]. Despite the significant effort that has been devoted to GWASs, most published studies have failed to identify such effective predictive biomarkers.

One of the most fundamental problems of gene-by-treatment interaction tests is their lack of sufficient statistical power. In general, these tests are based on ordinary regression models and have low statistical power compared to tests for detecting genetic main effects. The reason is that unreliability is compounded in the interaction term of the models since the tests assess the difference of magnitudes of gene effects between treatment groups rather than simply assessing the magnitudes of the gene effects themselves. A rule of thumb to detect interaction effects has been suggested, whereby the detection requires a sample size at least four times larger than that required for the detection of a main effect of comparable magnitude[4,5]. In addition, most GWASs strictly control the conservative genome-wide significance level ($p < 5 \times 10^{-8}$) for the interaction tests to adjust multiplicity. These conventional analysis strategies potentially set serious limits on the value of these studies' outcomes. While we in fact identified a

small number of useful biomarkers via these GWASs, the primary purpose of large-scale data analyses should be to effectively screen for genes that should be further investigated as candidate biomarkers in individualized precision medicine.

To overcome the lack of statistical power, an alternative effective multi-subgroup gene screening method[6] using a multidimensional semi-parametric hierarchical mixture model[7,8] has been developed by Matsui et al. (Figure 1). This method efficiently reveals the existence of predictive genes that are differently associated with outcomes between subgroups (treatment and control groups in a randomized clinical trial, for example), and prognostic genes that are similarly associated with outcomes irrespective of subgroups. To this end, the method eliminates two types of nuisance factors in association analysis results: (i) genes that are not associated with outcomes and (ii) random variation irrespective of association with outcomes. The method achieves these goals by using the hierarchical mixture model, and reveals the underlying effect size distribution of genes that are associated with outcomes. The denoised distribution itself could provide clues for exploring predictive biomarkers, optimizing individualized treatments, and understanding the biological mechanisms of diseases. One can demonstrate the existence of predictive/prognostic gene subgroups and their effect sizes from the distribution rather than by identifying individual genes using interaction tests with a conservative significance criterion. Furthermore, effect size estimates adjusted for overestimation error arising from association analyses, the so-called winner's curse phenomenon[9], can be obtained based on the distribution. The adjusted estimates would serve as fundamental information for developing appropriate therapeutic strategies. In addition, an efficient test can also be developed using the optimal discovery procedure (ODP)[10,11], which can provide an optimal ranking of genes as well as the most powerful

test for detecting disease-related genes with control of multiplicity, e.g., the false discovery rate (FDR). Note that the primary purpose of the newly developed method is to use a different principle than that underlying conventional gene-by-treatment interaction tests in order to provide an alternative strategy to more effectively identify predictive biomarkers, one that will overcome the issue of low statistical power and facilitate precision medicine research from a data analytic perspective.

In this article, we demonstrate the important strengths of these newly available tools by applying them to two large randomized clinical trials, the vitamin intervention stroke prevention (VISP) trial[12] and the SUCCESS-A trial, which study stroke and breast cancer, respectively, using genomic data (see Descriptions of GWAS Datasets and Section A in the Supplementary Notes) to detect single nucleotide polymorphisms (SNPs) that can be used to predict responses to therapeutics. We present the denoised effect size distributions of SNPs that are associated with medical outcomes, so as to assess the existence of predictive SNPs; disease-related SNPs detected by the ODP, along with their characteristics; and the results of a genomic investigation of these disease-related SNPs conducted using publicly available tools and databases, for the purposes of biological investigation and validation of the new method.

Materials and methods

Descriptions of GWAS datasets

The two datasets used in this analysis were deposited in the dbGaP database (available at https://www.ncbi.nlm.nih.gov/gap) and derived from the VISP trial (study accession number: phs000343.v3.p1) and the SUCCESS-A trial (study accession number: phs000547.v1.p1). For details, see Section A in the Supplementary Notes.

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 as well as nonfatal myocardial infarction or mortality. Subjects were randomly assigned to receive daily doses of the high-dose formulation (treatment group) or the low-dose formulation (control group). A total of 1533 subjects (760 assigned to the treatment group and 773 assigned to the control group) with 774670 SNPs passed a quality control filter (see Section B in the Supplementary Notes). In this study, we used this dataset to investigate SNPs associated with blood homocysteine levels which are strongly associated with the risk of stroke. We took as outcome the difference in blood homocysteine levels between baseline and the first post-baseline measurements, as in the study of Wakefield et al.[13]. Association tests were conducted using a linear regression model (see Section C in the Supplementary Notes).

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[14], improved progression-free survival, overall survival, and toxicity. Subjects were randomly assigned to chemotherapy with gemcitabine (treatment group) or without gemcitabine (control group). A total of 3289 subjects (1621 assigned to the treatment group and 1668 assigned to the control group) with 424121 SNPs passed the quality control filter (see Section B in the Supplementary Notes). In this study, we used this dataset to investigate SNPs associated with progression-free survival in breast cancer patients. Association

tests were conducted using a proportional hazards regression model (see Section C in the Supplementary Notes).

Multi-subgroup gene screening method

In this analysis, we used the efficient multi-subgroup gene screening method[6] (Figure 1) developed to overcome the problem of insufficient power to detect interaction effects (see Section D in the Supplementary Notes for details). Contrary to standard gene-by-treatment interaction tests using regression models with interaction terms between genes and treatments, this method first separates control and treatment groups. Then, association analyses using regression models without interaction terms are independently conducted for each group. As a result, summary statistics, i.e., estimated gene main effect sizes and their standard errors, are obtained for each group. After that, using the summary statistics, this method reveals the proportion of genes that are disease related and the underlying effect size distribution of disease-related genes across treatment and control groups via empirical Bayes estimation under the multidimensional hierarchical mixture model. Furthermore, posterior probabilities of association and effect size estimates adjusted for gene selection errors and overestimation for each SNP are obtained based on the estimated distribution. Finally, based on the optimal gene ranking and posterior probabilities of association for each gene derived from the fitted model, disease-related genes are detected by the ODP with control of FDR.

Genomic annotation

For biological investigation of detected SNPs and for validation of the new analysis method, we conducted genomic annotations on LD surrogates of SNPs detected by our analysis using the publicly available ENCODE (Encyclopedia of DNA Elements) tools. Genomic annotation on LD surrogates ($r^2 > 0.6$, 1000 Genomes Project EUR data) of

newly detected SNPs was conducted using HaploReg[15,16] v4.1 (available at http://archive.broadinstitute.org/mammals/haploreg/haploreg.php) and RegulomeDB[17] v1.1 (available at http://regulomedb.org/).

Results

Blood homocysteine levels

We conducted ordinary association analyses for control and treatment groups using the linear regression model and obtained effect size estimates (regression coefficients) and their standard errors for each SNP (Supplementary Fig. S1a). We then eliminated nuisance factors in these results using the hierarchical mixture model and obtained the effect size distribution of SNPs associated with homocysteine levels (Figures 2a and 2c).

The estimated distribution identifies the multi-subgroup SNPs that can be classified as possible prognostic or predictive biomarkers; peaks on the diagonal line in the first and third quadrants of the distribution correspond to prognostic markers, while others correspond to predictive markers. The proportion of disease-related SNPs was estimated to be 0.001. This suggests that 793 SNPs are associated with outcomes; note that the number of independently associated SNPs should be much smaller since some SNPs are in linkage disequilibrium (LD). The denoised distribution suggests that almost all SNPs have small effects on homocysteine levels regardless of control and treatment assignments, although the large peak is shifted slightly in the negative direction, corresponding to a decrease in homocysteine levels in the treatment group. The slight shift of the peak is due to the effect of high-dose administration of multivitamin tablets and is not related to any genetic properties.

In addition to the large peak, there is a small peak that deviates from the low-effect

area, suggesting the existence of SNPs that will predict the benefit of high-dose administration of multivitamin tablets. This peak has a small positive effect on low-dose administration (effect size of 0.2) but has a large negative effect on high-dose administration (effect size of -3.3). The location of the peak suggests the existence of strong interaction effects between particular SNPs and high-dose administration that drastically decrease homocysteine levels. The marginal distribution for control and treatment groups clearly shows this difference (Figure 2c). The deviant peak is shown for the high-dose formulation (treatment) group, while no corresponding peak exists for the low-dose formulation (control) group.

We next detected SNPs associated with homocysteine levels using the ODP (Table 1), based on the optimal ranking via posterior probabilities of association (Figure 3) obtained from the estimated effect size distribution. The ODP detected five independent SNPs (FDR<5%) that are associated with outcomes. Of these, rs12631354, rs2367209, and rs10017302 are newly detected in this analysis while others have previously been suggested as being associated with homocysteine levels[13]. In particular, rs12631354 and rs10017302 might significantly change the effect of the administration of multivitamin tablets. These have strong interaction effects with the high-dose administration of multivitamin tablets, drastically decreasing homocysteine levels. From the therapeutic point of view, these two SNPs might be useful predictive biomarkers.

The ODP also detected SNPs that have been suggested as being associated with homocysteine levels[13], as well as rs3736238, which has been previously reported as the most statistically significant SNP[13] and which reached genome-wide significance ($p < 5 \times 10^{-8}$) using the standard regression model with interaction terms (see Section E in the Supplementary Notes and Supplementary Fig. S2a). According to the adjusted

effect size estimates, rs3736238 shows the largest benefit for the high-dose administration group but demonstrates a harmful effect for the low-dose administration group (Table 1 and Supplementary Fig. S3a). Although the biological mechanisms of the harmful effect with low-dose administration are unclear, this result suggests the need for high-dose administration of multivitamin tablets in stroke patients. Another SNP, rs1739317, which has also been reported previously[13] and which reached a suggestive level ($p < 10^{-6}$) using the standard regression model (Supplementary Fig. S2a), was also detected. As with the newly detected SNPs, this SNP has a strong interaction effect with the high-dose administration that drastically decreases homocysteine levels and results in neither benefit nor harm with the low-dose administration. On the other hand, rs16893296 on chromosome 6, which has been suggested as being associated with homocysteine levels[13], was not detected by the ODP.

Breast cancer

As with the analysis of homocysteine levels, we conducted ordinary association analyses for control and treatment groups using the proportional hazards regression model and obtained estimated regression coefficients (log hazard ratios) and their standard errors for each SNP (Supplementary Fig. S1b). We then applied the hierarchical mixture model to the analysis results and obtained the effect size distribution of SNPs associated with progression-free survival in breast cancer patients (Figures 2b and 2d). The proportion of disease-related SNPs was estimated to be 0.002; although some of these SNPs are in LD, this finding suggests that 903 SNPs are associated with outcomes.

The estimated distribution is more spread out in the gemcitabine (treatment) group, although since the peak of the distribution is centered almost at 0, almost all

disease-related SNPs have small effects on progression-free survival in patients irrespective of treatment assignments. The difference in spread of effect size distributions between the control group and the treatment group is clearly revealed by the marginal distribution (Figure 2d**)**. The spread in the treatment group suggests the existence of "beneficial" and "harmful" SNPs, i.e., some SNPs might increase survival rate with gemcitabine administration while others might do the opposite. These estimates suggest that careful evaluations of effect sizes for each predictive SNP are necessary to develop appropriate therapeutic strategies.

We next applied the ODP to detect SNPs associated with progression-free survival in breast cancer patients based on the posterior probability of association (Figure 4). The ODP detected new six SNPs (FDR<5%) associated with progression-free survival in breast cancer patients (Table 1). Since the existence of harmful SNPs was suggested by the denoised distribution, we assessed the effect sizes of these SNPs adjusted for errors arising from the association analysis. The adjusted effect size estimates (Table 1 and Supplementary Fig. S3b) suggest that rs6712299 and rs17367673 have beneficial effects on progression-free survival with gemcitabine administration while rs4690351 and rs12449931 have harmful effects. Although further investigations are necessary, these SNPs might be useful predictive biomarkers to determine whether or not gemcitabine treatment should be conducted. On the other hand, rs12620133 and rs7311993 have beneficial effects in both the control and treatment groups. These SNPs might be used as prognostic biomarkers for developing risk-stratification systems.

External validation

We conducted genomic annotations on LD surrogates of SNPs detected by our analysis. For homocysteine level analysis, three SNPs (rs12631354, rs2367209, and rs10017302)

were used as queries, while for breast cancer analysis, six SNPs (rs12620133, rs6712299, rs4690351, rs7311993, rs12449931, and rs17367673) were used. Summaries of the results are presented in Supplementary Tables S1 to S9.

For blood homocysteine levels, we queried the three newly detected SNPs and obtained lists of LD surrogates with their characteristics (Supplementary Tables S1 to S3). As a high-LD surrogate of rs12631354, we identified rs4450813 ($r^2 = 1$ based on 1000 Genomes Project European population), which has an expression quantitative trait locus (eQTL) effect on the *RYK* gene in the liver[18] (Supplementary Table S1). This result seems biologically plausible because homocysteine is metabolized in the liver. For rs2367209, we identified the LD surrogate rs4679904 ($r^2 = 0.66$), which has an eQTL effect on the *AF038199* gene in liver tissue[19] and is associated with primary biliary cirrhosis[20] (Supplementary Table S2). The association between rs4679904 and primary biliary cirrhosis is strongly supported by existing evidence and is recorded in the National Human Genome Research Institute (NHGRI) GWAS catalog[21] (available at https://www.genome.gov/gwastudies/). This result suggests the existence of an actual regulatory variant, despite the fact that the effect size of rs2367209 on the high-dose administration of multivitamin tablets is comparably small (Table 1), and further investigation should be conducted. According to epigenomic information for another high-LD surrogate, rs1879797 ($r^2 = 0.84$), there is a cluster of active transcription start site/enhancer in digestive. This result might indicate the contribution of the variant to digestion of homocysteine. For rs10017302, the high-LD surrogates rs2126029 ($r^2 = 0.94$) and rs1460781 ($r^2 = 0.94$) were found, both of which have an eQTL effect for the *GPM6A* gene in peripheral blood monocytes[22] (Supplementary Table S3). Since the purpose of this study was to investigate associations with the transcriptome of

circulating monocytes, a key cell type involved in immunity-related diseases and atherosclerosis, the SNPs might consistently be associated with cardiovascular diseases. These results support the biological validity of the testing method under the hierarchical mixture models.

We also investigated SNPs that were associated with progression-free survival in breast cancer patients. As with the stroke trial, we queried six independent SNPs (Table 1) detected by our analysis and obtained lists of LD surrogates with their characteristics (Supplementary Tables S4 to S9). Some of these indicated the existence of biological mechanisms and might be truly associated with outcomes. Genomic annotation identified rs3821340 ($r^2 = 0.88$), which has eQTL effects for the *AC073464.7*, *CYP4F32P*, and *ZNF514* genes in the pancreas[23], as a surrogate of rs6712299 (Supplementary Table S5), as well as rs4690439 ($r^2 = 0.64$), which has an eQTL effect for the *WDR17* gene in the ovary[23], as a surrogate of rs4690351 (Supplementary Table S6). Although existing pharmacogenomic/pharmacogenetic genome-wide studies have identified SNPs that are associated with responses to gemcitabine[24–27] and their results are recorded in the NHGRI GWAS catalog[21], the newly detected SNPs did not match any records, partly because the previous studies focused on drug responses in pancreatic cancer, neutropenia, and/or leucopenia rather than breast cancer. Epigenomic information also suggests an association between some of the detected SNPs and breast cancer; enhancer activities in breast variant human mammary epithelial cells (vHMEC) exist for rs4690351, rs7311993, and rs17367673. Although there is still no strong evidence of an association with gemcitabine response and further validations are necessary, these results suggest the possibility of associations between these SNPs and breast cancer.

Discussion

In this analysis, we demonstrated an efficient multi-subgroup gene screening method using hierarchical mixture models and the ODP, with applications to molecular data from randomized clinical trials of stroke and breast cancer to detect predictive biomarkers. We found three new SNPs that were associated with blood homocysteine levels, and six new SNPs that were associated with progression-free survival in breast cancer patients. These SNPs have not been reported by existing GWASs.

This new method can more effectively detect predictive disease-related SNPs than conventional association tests that use regression models with interaction terms. For comparison, we conducted association tests based on standard regression models with interaction terms for each trial (see Section E in the Supplementary Notes). In the association analysis for homocysteine levels, only one SNP, rs3736238, reached a genome-wide significance level of $p < 5 \times 10^{-8}$, while rs16893296 and rs1739317 reached the suggestive level of $p < 10^{-6}$ (Supplementary Fig. S2a). Three other predictive SNPs detected by our analysis did not reach the suggestive level. On the other hand, rs16893296, which has been suggested as being associated with homocysteine levels[13], was not detected by our analysis. No signal peak corresponding to this SNP was found in the estimated effect size distribution, and the posterior probability of association was estimated as only 13%. Also, the adjusted effect size estimates suggest that this SNP does not have a large effect on homocysteine levels (Supplementary Fig. S3a). This result may indicate the possibility of a false positive in the existing study or a false negative in our analysis, and further validation is necessary. In the breast cancer trial, no SNPs reached the suggestive level (Supplementary Fig. S2b), while our analysis detected six independent SNPs consisting of four predictive

SNPs and two prognostic SNPs. We also conducted a comparison of the number of detected SNPs between the standard association tests and the new method under specified FDR levels (see Section E in the Supplementary Notes). The new method detected more SNPs associated with homocysteine levels than the standard method under the same FDR levels, and also, unlike the standard method, effectively detected several SNPs associated with progression-free survival in breast cancer patients (Supplementary Table S10). Note that categorization of SNPs as predictive markers or prognostic markers was conducted in a subjective manner. Basically, SNPs that have different effect sizes between control and treatment groups would be categorized as predictive markers, while others would be categorized as prognostic markers. However, a specific criterion to categorize SNPs will be subjectively determined (see Matsui et al.[6], for example). In addition, we assessed the performance of the ODP through a simulation study based on the two clinical trials (see Section F in the Supplementary Notes). The ODP detected larger numbers of significant SNPs with controlling FDR accurately, compared with the conventional methods (Supplementary Tables S11 and S12).

As demonstrated in this analysis, the denoised distribution can be used to explore the existence of predictive biomarkers and identify the best therapeutic strategies using a different approach than the ordinary gene identification scheme that uses gene-by-treatment interaction tests. For example, because the daily intake of high-dose multivitamin tablets has no serious harmful effects, such as increasing homocysteine levels, it can be administered to stroke patients according to the obtained distribution with no risk (Figures 2a and 2c). Also, in the low-dose formulation (control) group, all disease-related SNPs have effect sizes of nearly zero, corresponding to no impact on

homocysteine levels, and no peaks in the distribution deviate either positively or negatively (Figures 2c). This result means that low-dose administration does not maximize response or minimize side effects. From the therapeutic point of view, these findings suggest that high-dose multivitamin tablets should be administered to stroke patients irrespective of their genetic variations. On the other hand, the use of gemcitabine requires careful assessment of patients' genetic characteristics because the estimated distribution suggests that the drug might cause serious side effects in breast cancer patients with particular genetic variants and might decrease survival rates (Figures 2b and 2d). From a therapeutic point of view, these results indicate that individualized gemcitabine administration is necessary to improve survival rates and avoid side effects. Note that although this analysis strategy can reveal the existence of predictive/prognostic biomarkers as shown in the above examples, follow-up studies are necessary to definitively identify all of these markers, although some of them were successfully detected by our analysis using the ODP.

Although we demonstrated the two dimensional hierarchical mixture model to analyse molecular data consisting of two subgroups in this study, models with three or more dimensions can also be developed to explore higher order interactions. The existence of such interactions can be evaluated by the multidimensional models defining subgroups according to combinations of specific values of multiple variables.

The denoised distribution and the estimated proportion of disease-related SNPs can also be used for designing future medical genomics studies to identify predictive biomarkers. In particular, using the estimated results, we can obtain required sample sizes to find predictive biomarkers under a specified power and FDR[7]. Furthermore, the estimated proportions of disease-related SNPs suggest the existence of other underlying

markers that might have smaller effect sizes. Although efforts to obtain adequate numbers of samples are naturally crucial, it is equally important to develop more efficient association tests on gene-by-treatment interactions, as demonstrated in this analysis, as these tests would be serve as a realistic approach to discovering predictive markers.

Acknowledgments

This work was supported by CREST, Japan Science and Technology Agency (JPMJCR1412), the Practical Research for Innovative Cancer Control (17ck0106266 since 2017) from the Japan Agency for Medical Research and Development, and JSPS KAKENHI Grant Numbers JP16H06299 and JP17H01557.

Conflicts of interest

The authors declare no conflict of interest.

Supplementary information is available at European Journal of Human Genetics' website.

References

1 Daly AK. Genome-wide association studies in pharmacogenomics. *Nat Rev Genet* 2010; **11**: 241–246.

2 Motsinger-Reif AA, Jorgenson E, Relling MV *et al.* Genome-wide association studies in pharmacogenomics: successes and lessons. *Pharmacogenet Genomics* 2013; **23**: 383–394.

3 Wheeler HE, Maitland ML, Dolan ME, Cox NJ, Ratain MJ. Cancer pharmacogenomics: strategies and challenges. *Nat Rev Genet* 2012; **14**: 23–34.

4 Thomas D. Gene-environment-wide association studies: emerging approaches. *Nat Rev Genet* 2010; **11**: 259–72.

5 Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. *Int J Epidemiol* 1984; **13**: 356–365.

6 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.

7 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.

8 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.

9 Ferguson JP, Cho JH, Yang C, Zhao H. Empirical Bayes correction for the Winner's Curse in genetic association studies. *Genet Epidemiol* 2013; **37**: 60–68.

10 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.

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

12 Spence JD, Howard VJ, Chambless LE *et al.* Vitamin Intervention for Stroke Prevention (VISP) trial: rationale and design. *Neuroepidemiology* 2001; **20**: 16–

25.

13 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.

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

15 Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 2012; **40**: 930–934.

16 Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res* 2016; **44**: D877–D881.

17 Boyle AP, Hong EL, Hariharan M *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 2012; **22**: 1790–1797.

18 Innocenti F, Cooper GM, Stanaway IB *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet* 2011; **7**: e1002078.

19 Greenawalt DM, Dobrin R, Chudin E *et al.* A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res* 2011; **21**: 1008–1016.

20 Hirschfield GM, Liu X, Xu C *et al.* Primary biliary cirrhosis associated with HLA, IL12A, and IL12RB2 variants. *N Engl J Med* 2009; **360**: 2544–2555.

21 Welter D, MacArthur J, Morales J *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014; **42**: D1001–D1006.

22 Zeller T, Wild P, Szymczak S *et al.* Genetics and beyond – the transcriptome of human monocytes and disease susceptibility. *PLoS One* 2010; **5**: e10693.

23 GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015; **348**: 648–60.

24 Innocenti F, Owzar K, Cox NL *et al.* A genome-wide association study of overall survival in pancreatic cancer patients treated with gemcitabine in CALGB 80303. *Clin Cancer Res* 2012; **18**: 577–584.

25 Li L, Fridley BL, Kalari K *et al.* Discovery of genetic biomarkers contributing to variation in drug response of cytidine analogues using human lymphoblastoid cell lines. *BMC Genomics* 2014; **15**: 93.

26 Low SK, Chung S, Takahashi A *et al.* Genome-wide association study of chemotherapeutic agent-induced severe neutropenia/leucopenia for patients in Biobank Japan. *Cancer Sci* 2013; **104**: 1074–1082.

27 Kiyotani K, Uno S, Mushiroda T *et al.* A genome-wide association study identifies four genetic markers for hematological toxicities in cancer patients receiving gemcitabine therapy. *Pharmacogenet Genomics* 2012; **22**: 229–235.

Titles and legends to figures

Figure 1. Workflow of the gene screening method using multidimensional hierarchical mixture models. The top panels show the genotypic and phenotypic data of control and treatment groups. The sample sizes of control and treatment groups are denoted by N_0 and N_1, and M is the total number of SNPs. The middle panels show the summary statistics consisting of estimated effect sizes $b^{(0)}, b^{(1)}$ and their standard errors for the main effect of SNPs derived from association analyses using regression models without interaction terms. The bottom panels show the denoised effect size distribution of SNPs that are associated with outcomes, posterior probabilities of association, and adjusted effect size estimates for each SNP. Significant SNPs are detected by the optimal discovery procedure based on the distribution and posterior probabilities.

Figure 2. Effect size distributions of SNPs associated with blood homocysteine levels or progression-free survival in breast cancer patients. (a and b) Two-dimensional distributions of homocysteine levels (a) and breast cancer (b). The *x* axis represents the effect size for the control group and the *y* axis represents the effect size for the treatment group. (c and d) Marginal distributions of homocysteine levels (c) and breast cancer (d). The *x* axis represents the effect size and the *y* axis represents the probability density, and distributions marginalized by the control group and the treatment group are plotted.

Figure 3. Plots of posterior probabilities of association with blood homocysteine levels for each SNP based on the estimated effect size distributions. Posterior probabilities for each SNP (*y* axis) are plotted by chromosomal position (*x* axis) in a similar way as with a Manhattan plot. Red points with rsIDs denote the probabilities of significant SNPs detected by the ODP (FDR<5%) and are not in LD.

Figure 4. Plots of posterior probabilities of association with progression-free survival in breast cancer patients for each SNP based on the estimated effect size distributions. Posterior probabilities for each SNP (*y* axis) are plotted by chromosomal position (*x* axis) in a similar way as with a Manhattan plot. Red points with rsIDs denote the probabilities of significant SNPs detected by the ODP (FDR<5%) and are not in LD.

Table 1. New significant loci of blood homocysteine levels and breast cancer detected through the optimal discovery procedure.

					Effect size (95% CI)		Adjusted effect size (95% CI)		Posterior		
SNP	Chr.	Position	A_{12}	MAF	Control	Treatment	Control	Treatment	probability	Gene	Distance
Blood homocysteine levels											
rs12631354	3	134195045	CT	0.01	-0.07 (-1.41 to 1.28)	-3.90 (-5.24 to -2.56)	0.21 (0.19 to 0.42)	-3.30 (-3.46 to -1.55)	0.989	*RYK*	Intragenic
rs2367209	3	160681097	TG	0.27	0.14 (-0.28 to 0.56)	-0.99 (-1.40 to -0.58)	0.07 (-0.23 to 0.15)	-0.51 (-0.54 to -0.47)	0.871	*ARL14*	2.7 kb 3'
rs10017302	4	114563671	CT	0.02	0.52 (-0.85 to 1.89)	-3.63 (-4.89 to -2.37)	0.22 (0.20 to 0.69)	-3.30 (-3.63 to -1.17)	0.990	*UGT8*	35 kb 5'
rs1739317	6	24947873	TC	0.03	0.25 (-0.61 to 1.11)	-2.86 (-3.74 to -1.98)	0.22 (0.19 to 0.24)	-3.28 (-3.38 to -0.66)	>0.999	*FAM65B*	Intragenic
rs3736238	17	28881308	TC	0.01	1.88 (0.31 to 3.44)	-4.75 (-6.46 to -3.04)	1.35 (0.28 to 1.80)	-4.22 (-4.61 to -2.05)	0.982	*FLOT2*	Intragenic
Breast cancer											
rs12620133	2	3879585	AC	0.20	0.39 (0.14 to 0.64)	0.54 (0.28 to 0.80)	0.22 (0.09 to 0.36)	0.33 (0.20 to 0.47)	0.938	*DCDC2C*	32 kb 3'
rs6712299	2	95369055	CA	0.40	0.11 (-0.10 to 0.32)	0.58 (0.34 to 0.81)	0.07 (-0.07 to 0.22)	0.40 (0.28 to 0.50)	0.924	*KCNIP3*	Intragenic
rs4690351	4	176498757	GA	0.17	-0.05 (-0.35 to 0.25)	-0.69 (-0.95 to -0.44)	0.01 (-0.10 to 0.11)	-0.42 (-0.59 to -0.33)	0.963	*SPCS3*	167 kb 3’
rs7311993	12	11778652	AG	0.22	0.40 (0.14 to 0.66)	0.58 (0.33 to 0.84)	0.22 (0.08 to 0.36)	0.36 (0.25 to 0.51)	0.971	*ETV6*	Intragenic
rs12449931	17	79279996	GA	0.42	0.03 (-0.19 to 0.25)	-0.69 (-0.94 to -0.43)	0.03 (-0.07 to 0.14)	-0.44 (-0.59 to -0.33)	0.965	*RBFOX3*	Intragenic
rs17367673	19	46894238	AG	0.17	-0.14 (-0.41 to 0.14)	0.64 (0.39 to 0.89)	-0.06 (-0.18 to 0.08)	0.41 (0.31 to 0.53)	0.942	*ARHGAP35*	24 kb 5’

Effect size estimates and their 95% confidential intervals were obtained from association analyses under linear regression models for homocysteine levels and proportional hazards models for breast cancer. Effect size estimates for breast cancer correspond to log hazard ratios. Adjusted effect size estimates were calculated based on the denoised effect size distributions under the hierarchical mixture models. Positions based on hg38 were obtained from dbSNP build 141. Allele frequencies are based on 1000 Genomes Project European population and gene annotations are based on GENCODE version 13. Chr., chromosome. A_{12}, reference and alternate alleles.

Control group

M SNPs

N_0 individuals

	rs_1	rs_2	...	rs_M	Trait
$id_1^{(0)}$	1	0	...	0	-0.4
$id_2^{(0)}$	0	2	...	1	1.1
:	:	:		:	:
:	:	:		:	:
$id_{Nc}^{(0)}$	2	0	...	1	0.6

Association analysis

Summary statistics

M SNPs

	$b^{(0)}$	s.e.
rs_1	-0.2	0.3
rs_2	0.4	0.1
:	:	:
:	:	:
rs_M	0.3	0.5

Treatment group

M SNPs

N_1 individuals

	rs_1	rs_2	...	rs_M	Trait
$id_1^{(1)}$	2	1	...	0	-0.8
$id_2^{(1)}$	0	0	...	2	0.8
:	:	:		:	:
:	:	:		:	:
$id_{Nt}^{(1)}$	0	2	...	0	1.3

Association analysis

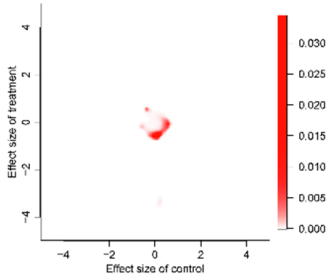
Summary statistics

M SNPs

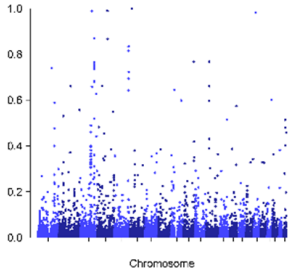
	$b^{(1)}$	s.e.
rs_1	-0.3	0.6
rs_2	0.3	0.2
:	:	:
:	:	:
rs_M	0.6	0.3

Empirical Bayes estimation

Denoised effect size distribution



Posterior probability of association

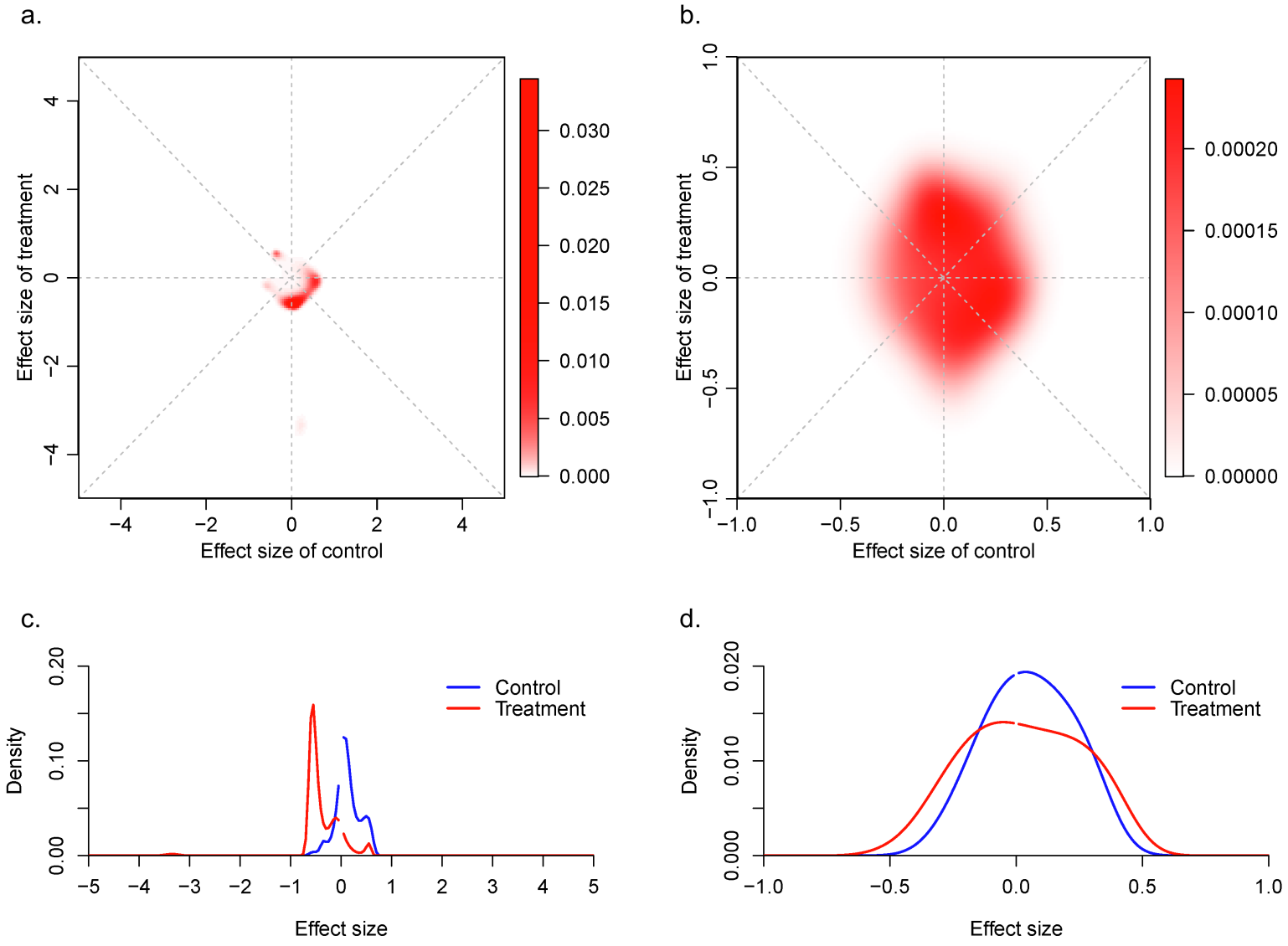


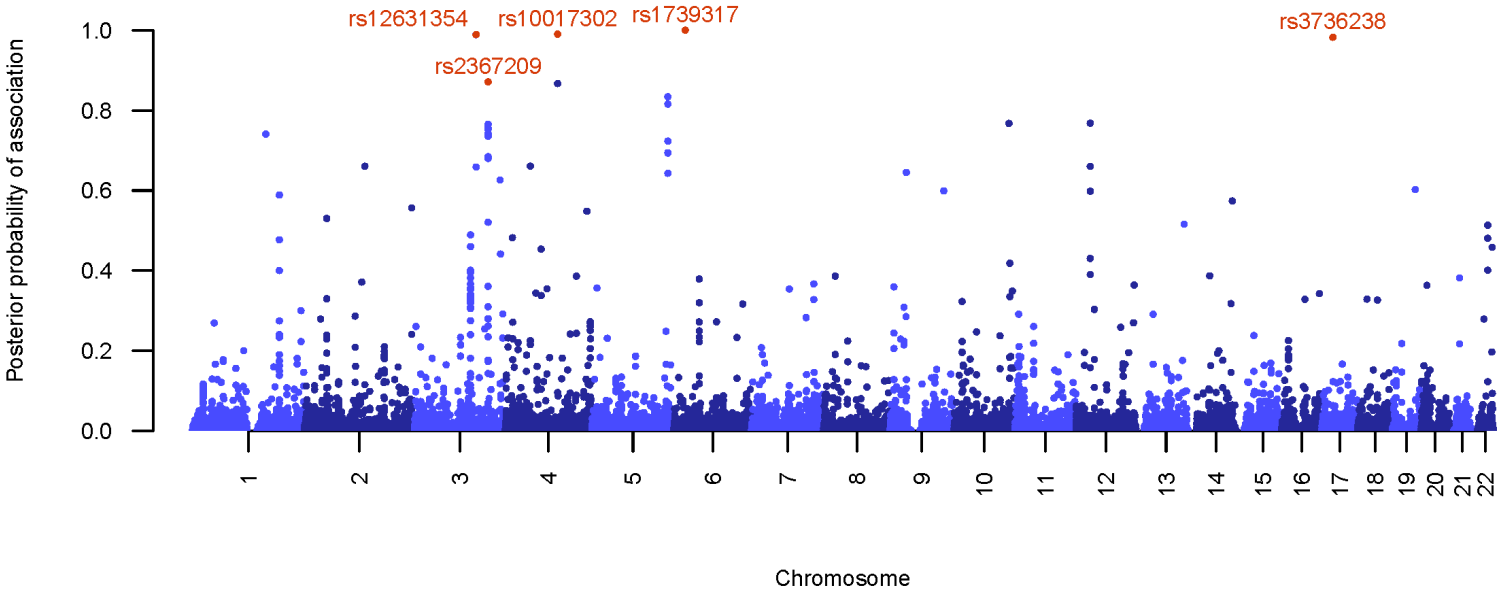
Adjusted effect size estimates

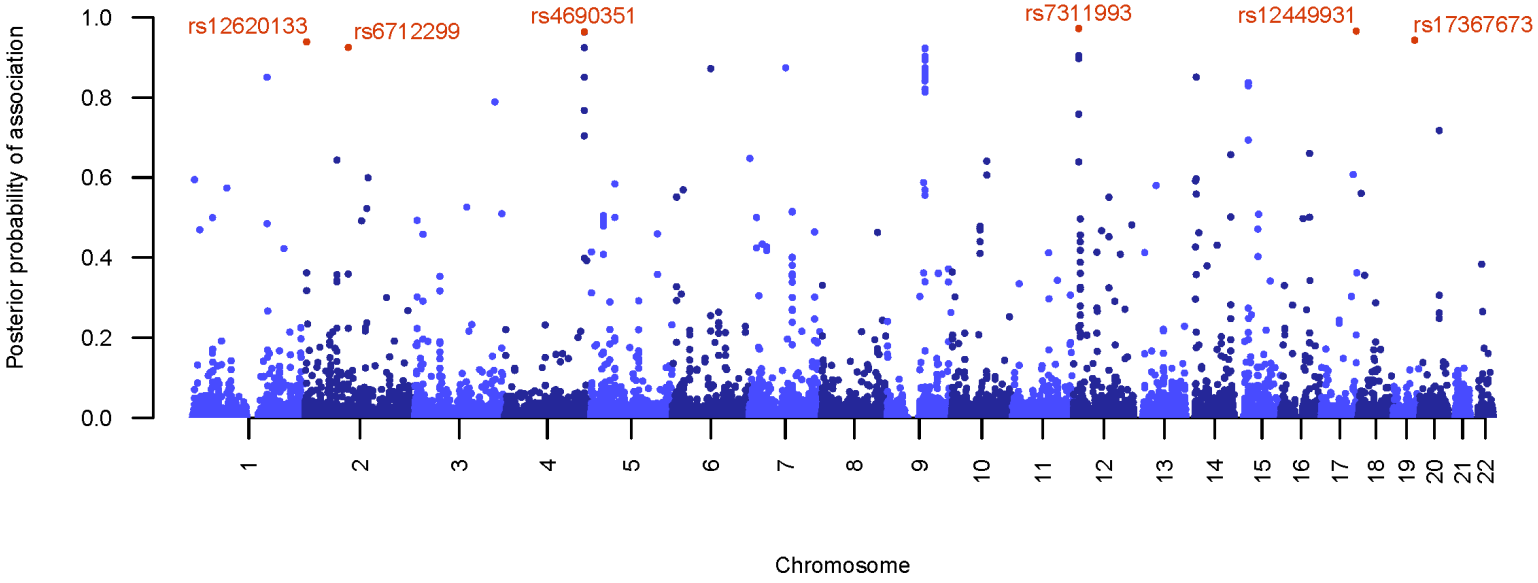
	$w^{(0)}$	$w^{(1)}$
rs_1	-0.1	-0.2
rs_2	0.3	0.2
:	:	:
:	:	:
rs_M	0.2	0.5

Optimal discovery procedure

Significant SNPs


a.
Effect size of treatment
Effect size of control
0.030
0.025
0.020
0.015
0.010
0.005
0.000
b.
Effect size of treatment
Effect size of control
0.00020
0.00015
0.00010
0.00005
0.00000
c.
Density
Effect size
Control
Treatment
d.
Density
Effect size
Control
Treatment


rs12631354
rs10017302
rs1739317
rs2367209
rs3736238
1.0
0.8
0.6
0.4
0.2
0.0
Posterior probability of association
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome


rs12620133
rs6712299
rs4690351
rs7311993
rs12449931
rs17367673
Posterior probability of association
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome