frontiers in Genetics

ORIGINAL RESEARCH
published: 24 April 2018
doi: 10.3389/fgene.2018.00115

Empirical Bayes Estimation of Semi-parametric Hierarchical Mixture Models for Unbiased Characterization of Polygenic Disease Architectures

*Jo Nishino[1,2], Yuta Kochi[2,3], Daichi Shigemizu[1,2,4,5], Mamoru Kato[2,6], Katsunori Ikari[2,7], Hidenori Ochi[2,8,9], Hisashi Noma[2,10], Kota Matsui[2,11], Takashi Morizono[5], Keith A. Boroevich[5], Tatsuhiko Tsunoda[1,2,5,12] and Shigeyuki Matsui[2,11,12]**

[1] Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo, Japan, [2] Core Research for Evolutionary Science and Technology (CREST), Japan Science and Technology Agency (JST), Tokyo, Japan, [3] Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan, [4] Division of Genomic Medicine, Medical Genome Center, National Center for Geriatrics and Gerontology, Obu, Japan, [5] Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan, [6] Department of Bioinformatics, National Cancer Center Research Institute, Tokyo, Japan, [7] Institute of Rheumatology, Tokyo Women's Medical University, Tokyo, Japan, [8] Division of Frontier Medical Science, Department of Gastroenterology and Metabolism, Programs for Biomedical Research Graduate School of Biomedical Science, Hiroshima University, Hiroshima, Japan, [9] Laboratory for Digestive Diseases, RIKEN Center for Integrative Medical Sciences, Hiroshima, Japan, [10] Department of Data Science, The Institute of Statistical Mathematics, Tokyo, Japan, [11] Department of Biostatistics, Nagoya University Graduate School of Medicine, Nagoya, Japan, [12] Risk Analysis Research Center, The Institute of Statistical Mathematics, Tokyo, Japan

Genome-wide association studies (GWAS) suggest that the genetic architecture of complex diseases consists of unexpectedly numerous variants with small effect sizes. However, the polygenic architectures of many diseases have not been well characterized due to lack of simple and fast methods for unbiased estimation of the underlying proportion of disease-associated variants and their effect-size distribution. Applying empirical Bayes estimation of semi-parametric hierarchical mixture models to GWAS summary statistics, we confirmed that schizophrenia was extremely polygenic [~40% of independent genome-wide SNPs are risk variants, most within odds ratio (OR = 1.03)], whereas rheumatoid arthritis was less polygenic (~4 to 8% risk variants, significant portion reaching OR = 1.05 to 1.1). For rheumatoid arthritis, stratified estimations revealed that expression quantitative loci in blood explained large genetic variance, and low- and high-frequency derived alleles were prone to be risk and protective, respectively, suggesting a predominance of deleterious-risk and advantageous-protective mutations. Despite genetic correlation, effect-size distributions for schizophrenia and bipolar disorder differed across allele frequency. These analyses distinguished disease polygenic architectures and provided clues for etiological differences in complex diseases.

Keywords: genome-wide association study (GWAS), polygenic disease architecture, polygenicity, effect-size distribution, semi-parametric hierarchical mixture model

OPEN ACCESS

Edited by:
Robert Klein, Icahn School of Medicine at Mount Sinai, United States

Reviewed by:
Heather E. Wheeler, Loyola University Chicago, United States
Gianluca Serafini, Dipartimento di Neuroscienze e Organi di Senso, Ospedale San Martino (IRCCS), Italy
Andrei Rodin, City of Hope National Medical Center, United States

***Correspondence:**
Shigeyuki Matsui
smatsui@med.nagoya-u.ac.jp

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *06 September 2017*
Accepted: *22 March 2018*
Published: *24 April 2018*

Citation:
Nishino J, Kochi Y, Shigemizu D, Kato M, Ikari K, Ochi H, Noma H, Matsui K, Morizono T, Boroevich KA, Tsunoda T and Matsui S (2018) Empirical Bayes Estimation of Semi-parametric Hierarchical Mixture Models for Unbiased Characterization of Polygenic Disease Architectures. Front. Genet. 9:115. doi: 10.3389/fgene.2018.00115

INTRODUCTION

Genome-wide association studies (GWAS) have identified numerous susceptibility variants for complex diseases (Welter et al., 2014). The sets of variants identified from GWAS, however, can generally explain only a small proportion of the heritability estimated from family studies, the so called "missing heritability" problem (Manolio et al., 2009). Much research has suggested that the variance explained by all SNPs in dense genotyping arrays, i.e., SNP heritability, often accounts for a large proportion of the family-based heritability (Lee et al., 2011, 2012, 2013; So et al., 2011b; Stahl et al., 2012; Ripke et al., 2013; Golan et al., 2014; Bulik-Sullivan et al., 2015; Palla and Dudbridge, 2015).

Quantitative evaluation of the polygenic architecture, in particular, the estimation of the proportion of disease-associated SNPs and their effect-size distribution, is essential to further determine the source of observed heritability (Wray et al., 2007; Park et al., 2010, 2011; Stahl et al., 2012; Agarwala et al., 2013; Chatterjee et al., 2013; Ripke et al., 2013). The estimation of these components also contributes to accurate power and sample size calculations of GWAS (Wray et al., 2007, 2012; Park et al., 2010; Yang et al., 2010; Ripke et al., 2013; Levinson et al., 2014) and estimation of the predictive capability of disease risks (Wray et al., 2007; Agarwala et al., 2013; Chatterjee et al., 2013).

However, we are still far from understanding the polygenic architecture of most complex diseases, because so far, there have been no feasible or fast methods that unbiasedly evaluate various polygenic architectures using the entire set of SNPs across the genome. Stahl et al. proposed estimating the proportion of disease-associated SNPs and the effect-size distribution using an approximate Bayesian polygenic analysis (Stahl et al., 2012). Its application, however, has been limited to few studies (Stahl et al., 2012; Ripke et al., 2013) because of the technical complexity and the excess computational burden of many simulations. On the other hand, some authors estimated the effect-size distribution based on a power evaluation for SNPs reaching genome-wide significance (Park et al., 2010, 2011; Chatterjee et al., 2013). This method, however, is to evaluate effect sizes only for those SNPs with relatively large effects, not all the disease-associated SNPs, requiring adjustment for the winner's curse (selection bias in using top significant SNPs) in the effect-size estimation.

To address the aforementioned limitations of existing methods, we propose an empirical Bayes estimation of semi-parametric hierarchical mixture models (SP-HMMs) (Matsui and Noma, 2011a,b) of GWAS summary statistics on effect sizes, such as estimated log-odds ratios to associate genotypes with disease susceptibility (see section Materials and Methods). This model decomposes GWAS summary data into signal and noise components and derives the proportion of disease-associated SNPs (non-null SNPs) and the distribution of their effect sizes (genotype log-odds ratios) as the signal component. To be more specific, mixture modeling refers to decomposing the underlying distribution of SNP-specific summary statistics into a non-null distribution for SNPs associated with disease occurrence, which corresponds to a signal component, and a null distribution for the remaining SNPs without association, which corresponds to a noise component, with a mixing probability or proportion of disease-associated SNPs, π. For the non-null distribution, semi-parametric hierarchical modeling incorporates standard asymptotic normality for summary statistics, while the true effect sizes follow a non-parametric prior distribution, g. With an expectation-maximization (EM) algorithm (Shen and Louis, 1999), we can estimate the prior probability π and distribution g using the data, i.e., empirical Bayes estimation. The empirical Bayes estimation of hierarchical mixture models is also applicable for SNP heritability estimation (So et al., 2011b) and adjustment for the winner's curse (Ferguson et al., 2013).

The features of our approach are summarized as follows: (1) the polygenic architecture for the entire set of SNPs, represented by π and g, can be flexibly and unbiasedly estimated, (2) it requires only summary data from GWAS (e.g., estimated log-odds ratios and standard error for individual SNPs are used), and (3) the estimation algorithm is easily implemented and fast. As such, the objective of the present study is to ascertain these features in evaluating the underlying the polygenic architecture of complex diseases, through its application to GWAS data from various diseases presumed to have distinct polygenic architecture in terms of several aspects.

Throughout this paper, we fit the SP-HMM to summary data from meta-/mega- GWAS analyses of rheumatoid arthritis (Okada et al., 2014), schizophrenia (Ripke et al., 2014), bipolar disorder (Sklar et al., 2011), and coronary artery disease [The (Coronary Artery Disease (C4D) Genetics Consortium., 2011; Schunkert et al., 2011)], to estimate the respective polygenic architectures and compare them across diseases. We also assess the liability-scale variance explained by SNPs, i.e., SNP heritability, based on this estimation. In order to obtain further insight into the underlying polygenic architectures, our approach can be applied to SNPs belonging to important functional categories, such as expression quantitative trait loci (eQTL), coding, non-synonymous, promoter, $5'$ or $3'$ UTR, enhancer, and DNase I hypersensitivity sites (Hindorff et al., 2009; Nicolae et al., 2010; Finucane et al., 2015; Gamazon et al., 2015). We focus on eQTLs, as gene expression levels have been increasingly recognized as notable endophenotypes or important mediators between genetic variations and disease phenotypes (Nicolae et al., 2010; Gusev et al., 2014; Gamazon et al., 2015; Zhu et al., 2016). Lastly, we also applied our method to GWAS data stratified by derived allele frequency (DAF), rather than minor allele frequency (MAF) (Park et al., 2011; Chan et al., 2014; Gorlov et al., 2015). A minor allele with low MAF can represent an allele with high DAF possibly under positive selection, as well as an allele that is maintained at low DAF through negative selection. Thus, our DAF-based analysis facilitates interpretation from the perspective of population genetics (Lachance, 2010), possibly contributing to further understanding of the genetic etiology for complex diseases.

RESULTS

We first confirmed the adequacy of our estimation method through unbiased estimation of the proportion of disease-associated SNPs, π, and their effect-size distribution, g, in simulation experiments (**Figure 1**; For all simulation settings, see Tables S1, S2, and Figures S1–S17 in Supplementary Material). The non-parametric estimation for g could flexibly capture various forms of underlying effect-size distributions. The execution time for the estimation using real data with one-hundred thousand SNPs was adequately fast; it completed in less than 5 min (Table S7).

For application to real GWAS datasets, we used publicly available summary statistics from large meta-/mega-GWAS for the four complex diseases (see Tables S3, S4 in Supplementary Material for details of the GWAS data). In associating each genotype with disease susceptibility, we defined the effect size as a log-odds ratio of the derived allele relative to the ancestral allele, denoted by β. We obtained an estimate of β and its variance estimate from the summary data. The ancestral/derived alleles for each SNP were determined from dbSNP.

Estimated Proportion of Disease-Associated SNPs and Effect-Size Distribution

To estimate the proportion of disease-associated SNPs, π, and the effect-size distribution, g, based on independent SNPs, we used two linkage disequilibrium (LD) pruned SNP sets: P-value-based and random-pruned sets. Note that we evaluated π and g with respect to the marginal effects of the nearly independent SNPs, as done by Stahl et al. (2012), rather than with respect to the effects of underlying causal variants themselves. The P-value-based method preferentially selected SNPs with stronger associations (hence more closely linked to causal variants) while using other GWAS data to correct for selection bias (see section Materials and Methods for details). The random-pruned method sampled SNPs randomly. In both methods, SNPs in LD ($r^2 > 0.1$) were removed. In a case where causal variants are in LD, only one would be retained in the final prune set, and thus, the estimates $\hat{\pi} \times$ (the number of SNPs in the SNP sets) using the pruned sets would give conservative estimates of the number of causal variants.

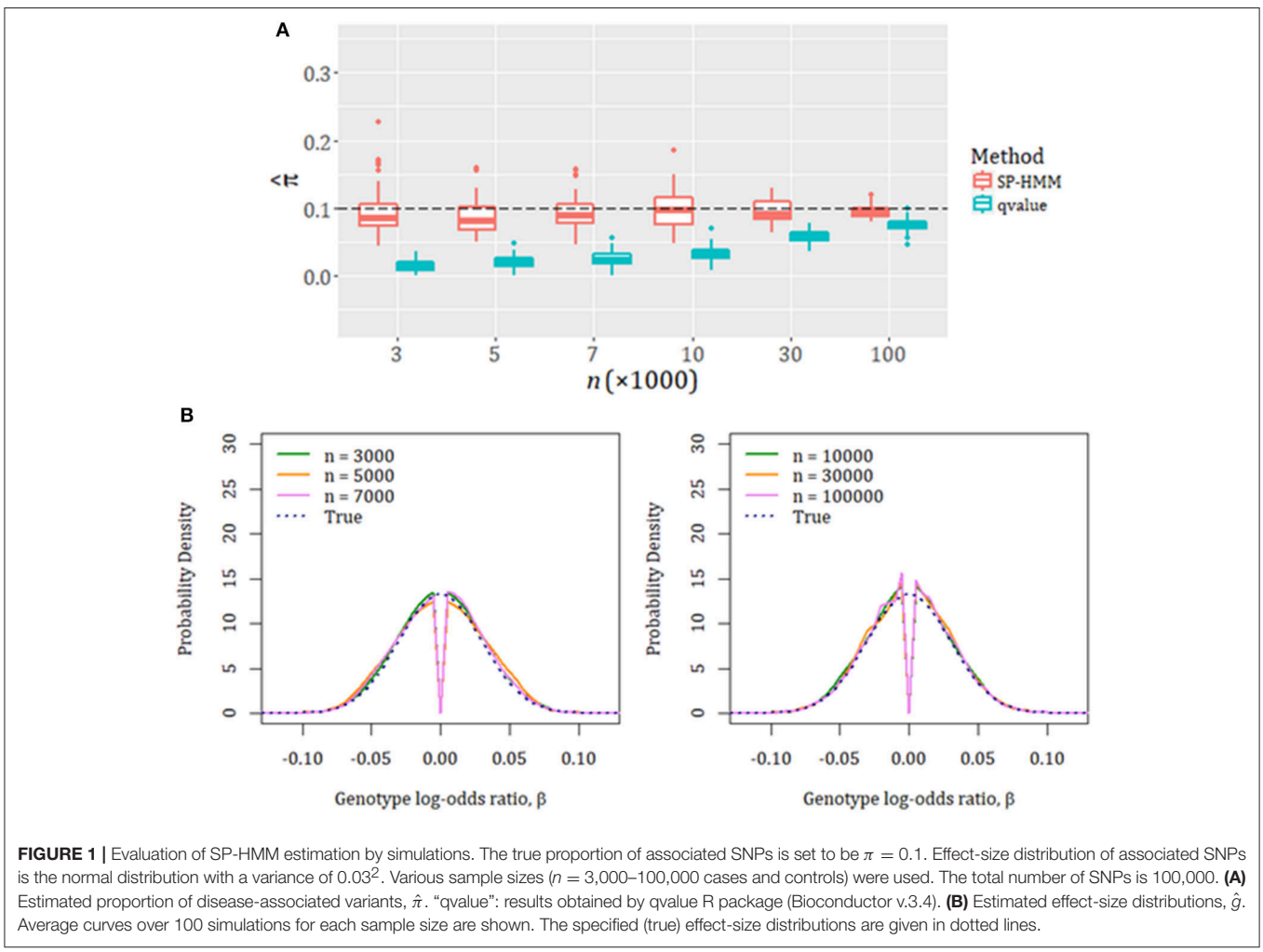

FIGURE 1 | Evaluation of SP-HMM estimation by simulations. The true proportion of associated SNPs is set to be $\pi = 0.1$. Effect-size distribution of associated SNPs is the normal distribution with a variance of 0.03^2. Various sample sizes (n = 3,000–100,000 cases and controls) were used. The total number of SNPs is 100,000. **(A)** Estimated proportion of disease-associated variants, $\hat{\pi}$. "qvalue": results obtained by qvalue R package (Bioconductor v.3.4). **(B)** Estimated effect-size distributions, $\hat{g}$. Average curves over 100 simulations for each sample size are shown. The specified (true) effect-size distributions are given in dotted lines.

We fit the SP-HMM to the *P*-value-based pruned SNP sets in each GWAS (**Table 1**; **Figure 2**). For rheumatoid arthritis, π was estimated as 3.6% for Asian and 8.1% for European populations, which were lower than the other diseases. The estimates of π were larger for two psychiatric diseases: 43.0% for schizophrenia and 39.6% for bipolar disorder. For coronary artery disease, using CARDIoGRAM and C4D data, π was estimated to be 15.9 and 26.1%, respectively.

With regard to the estimation of g, rheumatoid arthritis was shown to have a significant portion with larger effects, spanning to $|\beta| = 0.05$ (odds ratio of 0.95 or 1.05) or larger (**Figure 2**). It is noteworthy that, for rheumatoid arthritis, the proportion of positive effects was clearly larger than that of negative effects, indicating that the derived alleles are more likely to be risk alleles for the disease. Bipolar disorder was also estimated to have a distribution with relatively large effects. In contrast, schizophrenia and coronary artery disease was shown to have narrower distribution with very small effects. Schizophrenia was shown to have peaks around $|\beta| = 0.05$.

The estimates of π for the random-pruned SNP sets were similar to those for the *P*-value-based SNP sets for each GWAS (Table S5 in Supplementary Material). For the estimation of effect-size distribution, $\hat{g}$, the absolute effect size, $|\beta|$, tended to be slightly greater when using the *P*-value-based SNP sets than when using the random-pruned SNP set (Figure S19 in Supplementary Material).

Liability-Scale Variance Explained by the Pruned SNP Set

Using the estimates of the polygenic architecture (π and g), together with disease prevalence and allele frequencies, we could immediately evaluate the liability-scale variance, V, i.e., SNP heritability, explained by the pruned SNP set. Note that we evaluated V on the pruned SNP sets rather than on all SNPs on the GWAS chips. For evaluating V, the SP-HMM could directly model binary traits (i.e., disease occurrence) via log-odds ratios obtained from GWAS summary data.

Using the *P*-value-based pruned SNP sets, for rheumatoid arthritis, the estimates of V were 14.0% for Asian and 20.2% for European data (**Table 1**). Based on the estimated variance of 12% explained by the major histocompatibility complex (MHC) region (removed from the SNP set) and family based heritability of 55% (Supplementary Table 1 of Stahl et al., 2012), SNPs explained 47.3% (= (0.14 + 0.12)/0.55) and 58.2% (= (0.20 + 0.12)/0.55) of the family based heritability for the Asian and European populations, respectively, which were generally consistent with the previous estimate of 65% (Stahl et al., 2012). The estimates of V in schizophrenia and bipolar disorder were 40.2% and 50.0%, respectively, which were higher but almost within the range of previously reported estimates of 23-43% and 25-47%, respectively, for these diseases (Lee et al., 2012; Stahl et al., 2012; Ripke et al., 2013; Golan et al., 2014; Loh et al., 2015; Palla and Dudbridge, 2015). For cardiovascular disease, the estimates of V from the CARDIoGRAM and C4D data were 20.9 and 22.2%, respectively.

The estimates of V for the *P*-value-based pruned SNP sets (**Table 1**) were greater than those for the random-pruned SNP sets, but the differences were not substantial except for bipolar disorder (Table S5 in Supplementary Material).

TABLE 1 | Estimated proportions of disease-associated SNPs, $\hat{\pi}$, and liability-scale variance explained by SNPs, $\hat{V}$.

	$\hat{\pi}$ (SE[a]) (%)	$\hat{V}$ (SE[a]) (%)
Rheumatoid arthritis (Asian)	3.6 (1.8)	14.0 (1.8)
Rheumatoid arthritis (European)	8.1 (2.4)	20.2 (1.5)
Coronary artery disease (CARDIoGRAM)	15.9 (3.7)	20.9 (1.3)
Coronary artery disease (C4D)	26.1 (3.5)	22.2 (1.4)
Schizophrenia	43.0 (1.1)	40.2 (0.7)
Bipolar disorder	39.6 (2.2)	50.0 (1.9)

Estimates for the P-value-based SNP sets are shown.
For $\hat{V}$, disease prevalences are assumed to be 1% for rheumatoid arthritis and schizophrenia, 6% for coronary artery disease and 0.5% for bipolar disorder.
[a]*Estimated based on 100 parametric bootstrap samples based on the estimated SP-HMM.*

Stratified Estimation for eQTL/non-eQTL-SNPs

In order to gain insights into mediator effects of gene expression level, we fit the SP-HMM to "eQTL" SNPs, detected as cis-eQTLs using peripheral blood samples (Westra et al., 2013), and the remaining "non-eQTL"-SNPs, separately (**Figure 3**). All the SNPs in this analysis were selected to be nearly independent using a LD-pruning method based on LD ($r^2 > 0.1$) (see section Materials and Methods).

For rheumatoid arthritis in Asian and European populations, the proportions of disease-associated SNPs in the eQTL-SNPs were estimated to be larger than that in the non-eQTL-SNPs (**Figure 3**). In addition, the estimated effect-size distributions in terms of $\pi \times g$ (frequencies in the entire set including both null and non-null SNPs) in **Figure 3** indicated that there was a significant portion of SNPs with large effects, $|\beta| > 0.05$, for the eQTL-SNPs, but a small portion for the non-eQTL-SNPs, suggesting that the set of eQTL-SNPs included more components with distinctive large effects for rheumatoid arthritis. For the other diseases, there was a tendency for the frequencies of disease-associated SNPs in the set of eQTL-SNPs to be larger than those of the non-eQTL-SNPs.

We also estimated V for the eQTL-SNPs and non-eQTL-SNPs, separately (Table S6 in Supplementary Material). For rheumatoid arthritis, as expected from **Figure 3**, the per-SNP variance for the eQTL-SNPs was much larger than for the non-eQTL-SNPs. Interestingly, although eQTLs were defined using European samples (Westra et al., 2013), the enrichment of per-SNP variance (10.7-fold) in the eQTL-SNPs in the Asian population was larger than the 5.2-fold enrichment seen in the European population.

Estimation Across Derived Allele Frequencies

The effect size estimation of GWAS data stratified with the derived allele frequency (DAF) could provide another perspective

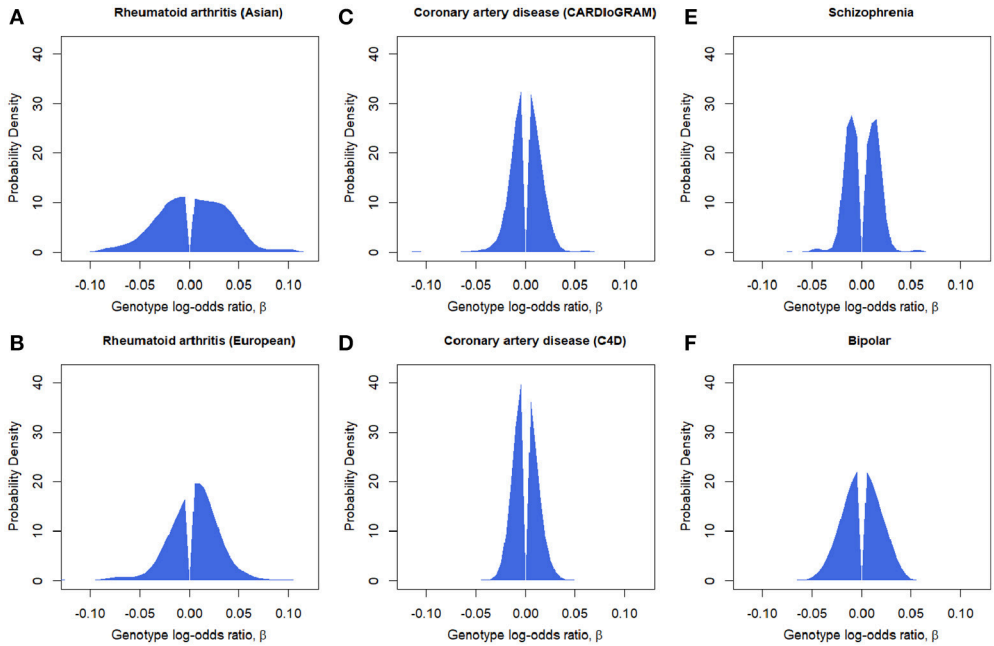

FIGURE 2 | Estimated effect-size distributions for disease-associated SNPs, $\hat{g}$. *P*-value-based pruned SNP sets are used. **(A)** Rheumatoid arthritis (Asian). **(B)** Rheumatoid arthritis (European). **(C)** Coronary artery disease (CARDIoGRAM). **(D)** Coronary artery disease (C4D). **(E)** Schizophrenia. **(F)** Bipolar disorder.

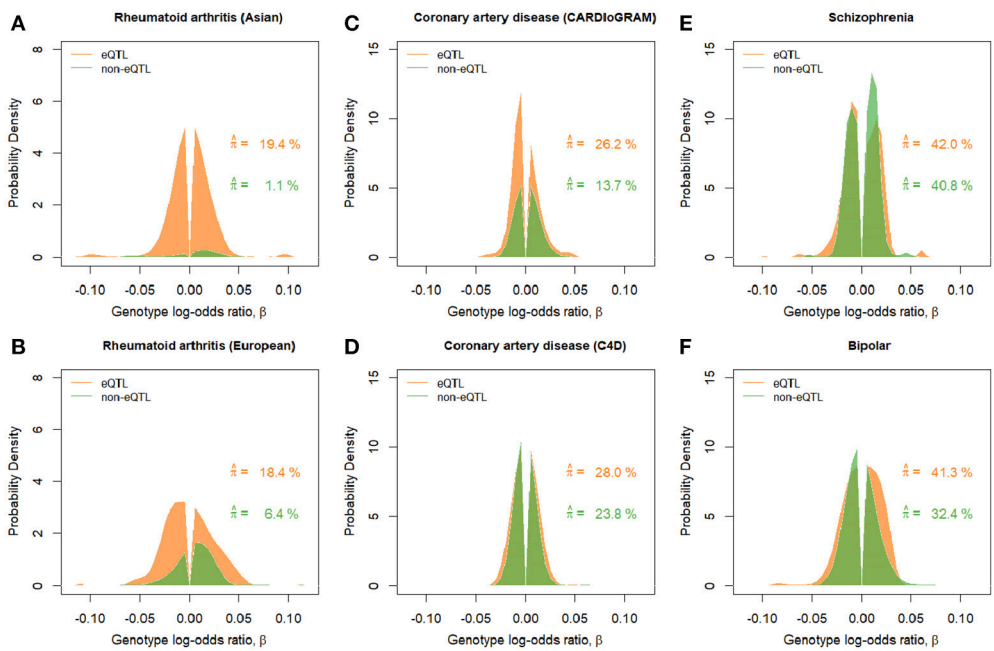

FIGURE 3 | Estimated effect size distributions for eQTL-SNPs and non-eQTL-SNPs, $\hat{\pi} \times \hat{g}$. Green and orange graphs show the results for the eQTL-SNP and non-eQTL-SNP sets, respectively. Estimated proportion of disease-associated SNPs, $\hat{\pi}$, correspond to the areas under the curves. **(A)** Rheumatoid arthritis (Asian). **(B)** Rheumatoid arthritis (European). **(C)** Coronary artery disease (CARDIoGRAM). **(D)** Coronary artery disease (C4D). **(E)** Schizophrenia. **(F)** Bipolar disorder.

on polygenic architecture, which facilitates assessment based on population genetics (see Discussion). We classified all SNPs into five equally-sized DAF bins and estimated the effect-size distribution for each bin. For rheumatoid arthritis, the estimated distributions across the DAF bins were similar between Asian and European data (**Figure 4**). We observed peaks at positive effects,

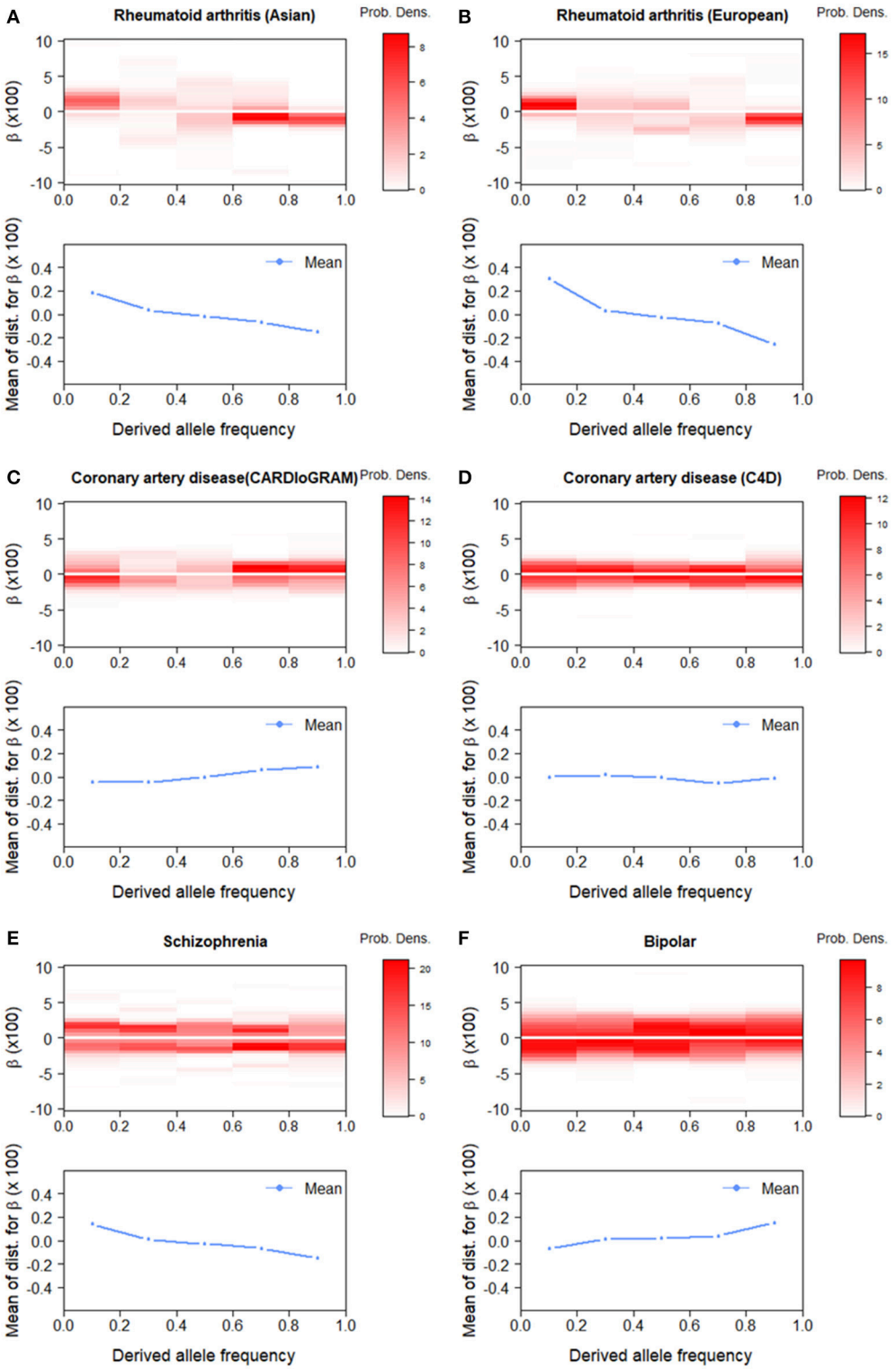

FIGURE 4 | Estimated effect-size distributions, $\hat{\pi} \times \hat{g}$, by derived allele frequency (DAF) bins. The upper panels (heatmap colors) for each GWAS results show $\hat{\pi} \times \hat{g}$. The lower panels show means of $\hat{\pi} \times \hat{g}$. **(A)** Rheumatoid arthritis (Asian). **(B)** Rheumatoid arthritis (European). **(C)** Coronary artery disease (CARDIoGRAM). **(D)** Coronary artery disease (C4D). **(E)** Schizophrenia. **(F)** Bipolar disorder.

i.e., $\beta > 0$, for lower DAF bins, especially for DAF ≤ 0.2, and at negative effects for higher DAF bins, especially for DAF > 0.8. This indicates that low-frequency-derived and high-frequency-derived alleles are prone to act as risk and protective variants for disease occurrence, respectively. For coronary artery diseases, there was no substantial difference in the estimated effect-size distribution among DAF bins, compared with rheumatoid arthritis. For schizophrenia and bipolar disorder, we observed opposite tendencies: for schizophrenia, positive and negative effects were over-represented, especially at DAF < 0.2 and DAF > 0.8, respectively, whereas, for bipolar disorder, negative and positive effects were over-represented at DAF ≤ 0.2 and DAF > 0.8, respectively.

DISCUSSION

We have developed a simple and fast method for unbiased estimation of the proportion of disease-associated variants and the effect-size distribution based on the empirical Bayes estimation of SP-HMM. As we hypothesized in the introduction, we observed that the SP-HMM provided new insights in evaluating polygenic models of complex diseases: The SP-HMM can effectively distinguish various polygenic architectures, including the degree of polygenicity and distributions of genotype log-odds ratio, across diseases, and can also provide various perspectives of the polygenic architecture based on important variant categories such as DAF and eQTL. To demonstrate the SP-HMM, we selected four diseases with relatively large GWAS (more than several thousand samples) to apply the model, as representatives of various types of complex diseases, i.e., autoimmune diseases, cardiovascular diseases, and psychiatric disorders. We summarized the findings obtained from the application of the SP-HMM to each disease together with the current understanding of their respective genetic architectures in the literature and discuss the similarities and differences in the results across the diseases, and generalize these results to other complex diseases and discuss limitation of the method, accounting for eQTL/no-eQTL and DAF category.

Schizophrenia has long been suspected to be polygenic (Gottesman and Shields, 1972; Purcell et al., 2009). The estimated SNP heritability of the disease, 23–43%, suggests that common and weak-effect SNPs not reaching significance can explain the moderate to high degree of family-based heritability (Lee et al., 2012, 2013; Ripke et al., 2013; Golan et al., 2014; Loh et al., 2015; Palla and Dudbridge, 2015). An approximate Bayesian polygenic analysis (ABPA; Stahl et al., 2012) estimated that 8,300 independent SNPs contributed to the genetic basis of schizophrenia and that genotypic relative risks for schizophrenia were relatively small compared with the other complex diseases (Figure 3 in Ripke et al., 2013). Through a simulation-based method, schizophrenia has been shown to be extremely highly polygenic compared with the other diseases examined and to have more than 20,000 causal variants (Loh et al., 2015). The extremely high polygenicity has been also confirmed by the observation that local SNP heritability estimates in independent LD blocks for schizophrenia were the most ubiquitously distributed among six complex diseases (Shi et al., 2016). Here, using the SP-HMM, in schizophrenia π was estimated to be ∼40% of disease-associated variants of independent SNPs in the genome (**Table 1** and Table S5 in Supplementary Material). This suggests at least ∼40,000 causal variants exist in the genome. The effect-sizes of the variants were clearly estimated to be very small for the most part, i.e., within $|\beta| = 0.03$, but larger than $|\beta| = 0.05$ for a small number of variants (**Figure 2** and Figure S19 in Supplementary Material). The clear-cut visualization of the effect-sizes for schizophrenia is new finding, as it is for the three other diseases examined.

Bipolar has been estimated to have 25–47% SNP-heritability (Lee et al., 2013; Golan et al., 2014). Despite limited significant disease-associated variants identified (Sklar et al., 2011), the estimates show that common and weak-effect SNPs can explain moderate to high degree of family-based heritability. The SP-HMM estimated ∼40% of independent SNPs in the genome to be disease-associated variants and the effect-sizes of the variants were estimated to be small for the most part, but ranging to around or more than $|\beta| = 0.05$ (**Table 1** and **Figure 2**, and, Table S5 and Figure S19 in Supplementary Material).

For rheumatoid arthritis, the SNP heritability has been estimated to be relatively small, 13–18% (Stahl et al., 2012; Palla and Dudbridge, 2015). Rheumatoid arthritis has been identified as a disease for which the majority of the SNP heritability can be explained a small percentage of the genome (Shi et al., 2016). Our estimates of π, 3.6% for Asians and 8.1% for Europeans, were generally consistent with the previous estimates of 2.7% by ABPA (Stahl et al., 2012) and 5.4% (Palla and Dudbridge, 2015) for Europeans, and a significant portion of the estimate for g ranged up to a $|\beta| = 0.05$ and extended so far as 0.1 (**Table 1** and **Figure 2**). In the rheumatoid arthritis stratification analysis based on eQTLs, we observed a high enrichment of per-SNP variance due to eQTLs determined by peripheral blood samples (Table S6 in Supplementary Material), similar to the enrichment on per-SNP variance by blood-specific DNaseI hypersensitivity sites (DHS) (Gusev et al., 2014), which were also strongly associated with expression variation (Degner et al., 2012). As peripheral blood samples include multiple types of leukocytes, the eQTLs have the potential to control immune-related gene expressions that are associated with the occurrence of rheumatoid arthritis. Although eQTLs were defined using European samples (Westra et al., 2013), the enrichment of 10.7-fold in the Asian population was larger than the 5.7-fold enrichment observed in the European population. The same tendency has been observed for the validated 100 non-MHC SNPs (Extended Data Figure 5 in Okada et al., 2014). This might be explained by non-eQTL-SNPs with large effects, such as non-synonymous SNPs in genes PTPN22 (R620W) and TYK2 (P1104A), which exist in Europeans but are absent or exist to a lesser degree in Asian populations. Some eQTL-SNPs were estimated to have large effect size $|\beta| > 0.05$ (**Figure 3**) in rheumatoid arthritis.

For coronary artery disease, the SNP heritability has been estimated to be 33–48% (Stahl et al., 2012; Golan et al., 2014). The SP-HMM estimated that π for C4D was larger (estimates ∼26% in the both of *P*-value-based and randomly-pruned sets)

than that for CARDIoGRAM (estimates of 15.9 % and 23.5 % in the *P*-value-based and randomly-pruned sets, respectively). Since SNPs of C4D were pruned by using LD structure of European ancestry (see section Materials and Methods), LD remaining in Asian SNPs, possibly linked with one causal variant, might increase the estimated proportion of disease-associated SNPs.

Regarding the similarity and differences among the four diseases, we identified a common feature across the four complex diseases for which the genetic basis consists of enormous variants (more than several thousand independent risk variants; **Table 1** and Table S5 in Supplementary Material) with very small effects (majority of genotypic OR for risk alleles are within 1.05; **Figure 2** and Figure S19 in Supplementary Material). The recently proposed "omnigenic" model hypothesizes that "gene regulatory networks are sufficiently interconnected such that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes and that most heritability can be explained by effects on genes outside core pathways" (Boyle et al., 2017). This model explains the existence of such numerous risk-variants with small effects. Next, the SPHMM showed that polygenicity was estimated to widely vary among the four diseases, e.g., schizophrenia was extremely polygenic (∼40% of independent genome-wide SNPs were risk variants, most within odds ratio OR = 1.03), whereas rheumatoid arthritis was less polygenic (∼4 to 8% risk variants, with a significant portion reaching OR between 1.05 and 1.1). The fewer but relatively large-effect variants would be the reason why the number of GWAS hits for rheumatoid arthritis (∼100 significant SNPs) is comparable to that for schizophrenia despite of small SNP heritability compared with schizophrenia. In fact, the effect sizes of validated variants for rheumatoid arthritis were generally larger than those for schizophrenia (Okada et al., 2014; Ripke et al., 2014). Our estimate of g means that the effect sizes of variants that will be detected in future should also be large for rheumatoid arthritis relative to other complex diseases.

Using DAF-stratified analysis for rheumatoid arthritis, we estimated more risk/protective derived alleles in low/high DAF (**Figure 4**). Simple models based on a theory of population genetics for DAF (Sawyer and Hartl, 1992) (see Figure S20 in Supplementary Material) could help interpret results from the DAF analysis, and thus provide another perspective on the differences among diseases (see Supplementary Note in Supplementary Material for details). Among such models, the "deleterious-risk and advantageous-protective mutation" model with weak selection was best fitted for rheumatoid arthritis (Figure S21 in Supplementary Material). Because most of the risk genes for rheumatoid arthritis are implicated in immune system regulation (Okada et al., 2014), these low- and high-derived alleles would tend to skew an individual's immune function toward either deleterious or beneficial directions. Meanwhile, this skewing may result in breaking the balance between immunity and tolerance, leading to rheumatoid arthritis.

Although some authors have reported that bipolar disorder and schizophrenia share a large amount of genetic factors (Purcell et al., 2009; Lee et al., 2013), we observed opposite tendencies in the genetic architecture for these diseases: risk (protective) and protective (risk) derived alleles were over-represented, especially at DAF $\leq$ 0.2 and DAF $>$ 0.8 for schizophrenia (bipolar disorder) (**Figure 4**). This paradoxical result was consistent with a previous report that, among low minor allele frequency (1–5%) SNPs, the R/P ratio (ratio of the number of detected variants with risk in minor allele to those with protective effect) for schizophrenia was significantly larger than one, while for bipolar disorder it was less than one (see Table 1 in Chan et al., 2014). Again, applying the same population genetics models, it was found that both the "deleterious-risk and advantageous-protective mutation" and "deleterious-risk mutation" models were better fitted for schizophrenia, whereas the "advantageous-risk and deleterious-protective mutation" model was the best fitted for bipolar disorder (Figure S21 in Supplementary Material). Recently, genetic correlations between creativity and both schizophrenia and bipolar disorder were reported, but they were much stronger for bipolar disorder (Keller and Visscher, 2015; Power et al., 2015). There is possibly some relationship between creativity and over-represented (positively selected) risk mutations at high DAF in bipolar since creativity is an important aspect for humans. In this way, the SP-HMM might provide a clue for resolving the shared and specific genetic etiologies between the two genetically related diseases.

The SP-HMM can also provide posterior effect-size estimates of individual SNPs based on the estimated genetic architecture, $\hat{\pi}$ and $\hat{g}$ (Stephens and Balding, 2009; Matsui and Noma, 2011b). To evaluate individual eQTL-SNPs, we used the estimated genetic architecture as the prior and listed the top SNPs with larger posterior means of effect size, $|\beta| > 0.05$ (Data Sheet 1 in Supplementary Material). As this list includes eQTLs such as RNASET2 and ADO, which have not been previously linked to rheumatoid arthritis (Okada et al., 2014), this approach might be effective for identifying disease associated eQTL-SNPs. For the other diseases, enrichments of per-SNP variance due to the eQTLs in peripheral blood cells were also observed. Since the eQTL-SNPs are associated with immune-related gene expression, these observations were consistent with the fact that coronary artery disease is a chronic inflammatory disorder and previous reports of the genetic overlap between immune diseases and schizophrenia (Stringer et al., 2014). However, it should be noted that precise estimation of the eQTL effects in these diseases requires additional eQTL data covering all the tissues and cells related to the diseases.

Although we only examined four complex diseases so far, the feature of enormous risk variants with very small effect could be generalized to almost all other complex diseases based on our experience analyzing several other diseases. It should be noted that polygenicity should generally differ among complex diseases even among those that belong to the same categories, i.e., psychiatric disorders. Specifically, whether the GWAS of a particular disease with a realizable sample size would successfully detect disease-associated variants largely depends on the existence of variants with relatively large effects, e.g., genotypic odds ratio >1.05, or >1.10. The number of such variants would vary largely between complex diseases.

The limitation of our method is that SP-HMM evaluates π and g with respect to the marginal effects of SNPs rather than with respect to the effects of underlying causal variants themselves. Nevertheless, the results of the SP-HMM estimation reflect the effects of the causal variants themselves through linkage disequilibrium.

Lastly, the SP-HMM and empirical Bayes method, which can provide fine characterization of genetic architecture, can also contribute to accurate power analysis of GWAS (Park et al., 2010; Ripke et al., 2013) and estimation of the predictive capability of disease risk (Chatterjee et al., 2013). The SP-HMM can also be extended to multi-dimensional settings, e.g., for quantification of sex in the genetic architecture of a disease, or the (antagonistic) pleiotropic genetic architecture in multiple diseases. This kind of multi-dimensional analysis would be novel and could provide new perspectives on multi-dimensional genetic effects, e.g., through a two-dimensional visualization of effect-size distributions for schizophrenia and bipolar diseases. Such analyses will be applied in future reports.

MATERIALS AND METHODS

Semi-parametric Hierarchical Mixture Model (SP-HMM)

We defined the effect size, β_j, for the j-th SNP of the total m SNPs as the genotype log-odds ratio under the additive allele dosage model. We considered the dosage of "derived mutant" alleles. Namely, the genotypes AA, Aa, and aa in each SNP had dosages $x_j = 0$, 1, and 2, respectively, where a was the derived and A was the ancestral allele. $Y_j = \hat{\beta}_j$ was an estimate of log-odds ratio for the j-th SNP (e.g., the standard maximum likelihood estimate). For each Y_j, we assumed a mixture structure with two components, null and non-null SNPs, in terms of association with disease susceptibility. To be specific,

$$f_j\left(y_j\right) = (1-\pi) f_{0j}\left(y_j\right) + \pi f_{1j}\left(y_j\right), \tag{1}$$

where f_{0j} and f_{1j} are the probability densities for null and non-null SNPs, respectively, and π is the prior probability of being non-null. For null SNPs, we specified $y_j \sim f_{0j}\left(y_j\right) = N\left(0, \hat{V}_{\hat{\beta}_j}\right)$ based on the asymptotic distribution of $\hat{\beta}_j$, where $\hat{V}_{\hat{\beta}_j}$ is an empirical variance estimate of $\hat{\beta}_j$ (e.g., the standard Wald-type variance estimate for $\hat{\beta}_j$). For non-null SNPs, we assumed the hierarchical structure: $y_j|\beta_j \sim f_{1j}\left(y_j|\beta_j\right) = N\left(\beta_j, \hat{V}_{\hat{\beta}_j}\right)$ and $\beta_j \sim g$, where the prior effect-size distribution g was unspecified. In this model, the standard asymptotic normality was assumed for $\hat{\beta}_j$ at the individual SNP level, while its true effect size β_j followed a non-parametric prior distribution g, forming a semi-parametric hierarchical mixture model (SP-HMM) (Matsui and Noma, 2011a,b). The assumption that each y_j is mutually independent would be reasonable for a set of LD-pruned SNPs.

Empirical Bayes Estimation

We estimated the priors, π and g, in the SP-HMM based on the data by applying an expectation–maximization (EM) algorithm, called the smoothing-and-roughening algorithm (Shen and Louis, 1999), to incorporate the non-parametric prior distribution g(Matsui and Noma, 2011a,b). The non-parametric estimate of g was supported by fixed discrete mass points $\boldsymbol{p} = (p_1, p_2, ..., p_B)$ at a series of nonzero points $\boldsymbol{b} = (b_1, b_2, ..., b_B)$ $(b_1 < b_2 < \cdots < b_B)$. We specified a wide range for the mass points, such as $b_1 = -0.3$ to $b_B = 0.3$ (0.74 to 1.35 in odds ratio), to support the effect-size distributions in many complex diseases. We set the number grid points as 120, such that $\boldsymbol{b} = (-0.300, -0.295, \ldots, -0.005, 0.005, \ldots, 0.295, 0.300)$. The initial value of π, π_{init}, and the initial distribution of g, g_{init}, were determined sequentially. Setting g to be uniformly distributed (i.e., $p_i = 1/B$ for all i), the EM procedures for candidate initial values, $\pi = 0.1$, 0.2, ..., or 0.9, were run 200 times and the value of estimated π with maximum likelihood was selected as π_{init}. Then setting g to be uniformly distributed again, we got g_{init} by the EM procedure with fixed $\pi = \pi_{init}$(the EM iterations were stopped when the relative change of π in one iteration was $< 0.005\%$ or after 200 iterations). The final EM procedure set $g = g_{init}$ and $\pi = \pi_{init}$, and was stopped when the relative changes in the estimate of π in one iteration was < 0.005 % or 2000 iterations was reached. We applied a parametric bootstrap method based on the estimated SP-HMM to estimate standard errors of the estimate for π.

Liability-Scale Variance Explained by SNPs

As shown by So et al. (2011a), the log odds ratio, β_j, together with the allele frequency and the disease prevalence, can be transformed to the variance explained by the j-th SNP, denoted as v_j, in the liability threshold model. In the liability threshold model, we assumed that an underlying liability to disease follows a normal distribution and individuals that exceeded a threshold of liability, T, were affected with the disease. Individuals with the genotypes of AA, Aa, and aa at the j-th locus had liability distributions with different means, but the same residual variance. We let p_j be the derived allele frequency and h_{j,x_j} be the frequency of genotype x_j ($x_j = 0$, 1, 2) in the general population. Assuming Hardy-Weinberg equilibrium in the population, the genotype frequencies are given by $h_{j,0} = (1-p_j)^2$, $h_{j,1} = 2p_j\left(1-p_j\right)$, *and* $h_{j,2} = {p_j}^2$. Using the overall mean liability, μ_{all}, and the mean liabilities of genotype x_j, μ_{j,x_j}, the variance explained by j-th SNP is given by

$$v_j^* = \sum\nolimits_{x_j=0}^{2} h_{j,x_j}\left(\mu_{j,x_j} - \mu_{all}\right)^2. \tag{2}$$

For evaluating μ_{j,x_j}, we used the penetrance of genotype x_j, denoted by $\varphi_{j,x_j} = 1/(1 + e^{-\alpha_j - \beta_j x_j})$ under the additive allele dosage model, where α_j was determined under the constraint involving the disease prevalence K, $K = \sum_{x_j=0}^{2} h_{j,x_j}\,\varphi_{j,x_j}$. Assuming that the residual variance of each genotype was 1, the mean liability of each genotype was given by

$$\Phi^{-1}\left(1-\varphi_{j,x_j}\right) = T - \mu_{j,x_j} \text{ for } \mathrm{x_j} = 0, 1, \text{ and } 2, \tag{3}$$

from which we obtained values of μ_{j,x_j}, where Φ was the cumulative distribution function of the standard normal

distribution. Of note, one of the mean liabilities of genotypes can be set as an arbitrary value, as it does not affect the variance estimate. Finally, v_j was obtained by $v_j = v_j^*/(1 + v_j^*)$. This corresponded to the variance under the standard liability threshold model with the unit total variance of liability, as is assumed in heritability estimation (Falconer, 1965; Lee et al., 2011).

We estimated the distribution of v_j for non-null effects using the estimated effect-size distribution $\hat{g}$, together with using allele frequencies and the prevalences. The allele frequencies were retrieved from the 1000 Genomes Project Phase 3 (The 1000 Genomes Project Consortium., 2015) and the prevalences previously assumed in estimating SNP heritability were used (Stahl et al., 2012; Golan et al., 2014). Then, the point estimate of v_j, $\hat{v}_j$, was gained as the product of the estimate $\hat{\pi}$ and the mean of the estimated distribution of v_j for non-null effects. The total liability-scale variance, V, explained by the pruned SNP sets, was then estimated as a simple sum of $\hat{v}_j$ over all SNPs in the sets.

GWAS Data Analysis

The six sets of GWAS summary statistics that we used were available online. The characteristics of individual GWASs are shown in Tables S1, S2 in Supplementary Material. For rheumatoid arthritis, the MHC region (chromosome 6, 25–35 Mb) was removed. The derived/ancestral states of alleles were determined by using dbSNP.

We used two kinds of pruned SNP sets, *P*-value-based and random-pruned sets, in the non-stratified SP-HMM analysis (**Table 1** and **Figure 2**). To gain the *P*-value-based pruned set for a GWAS, we began by selecting the most strongly associated SNP, i.e., the SNP with the lowest P value, in a reference GWAS as a SNP of the pruned set, and all other SNPs in LD ($r^2 > 0.1$) with the selected SNP were removed. The process was repeated until no SNPs remained. LD information was retrieved from the HapMap data base (HapMap phases I+II+III, release 27) (International HapMap 3 Consortium, 2010). In selecting SNPs with strong associations for Asian rheumatoid arthritis GWAS, European rheumatoid arthritis GWAS data were used as a reference for association, and vice versa. For coronary artery disease, the data of two GWAS, CARDIoGRAM and C4D, were used reciprocally. For the two genetically correlated diseases, schizophrenia and bipolar disease, the data of two GWAS for the two diseases were used reciprocally. For the random-pruned sets, we included SNPs randomly, irrespective of degrees of association, i.e., *P*-values in the reference GWAS data, such that no SNPs in the set were in $r^2 > 0.1$.

For stratified analysis by eQTL/non-eQTL-SNPs, we defined an "eQTL SNP" as a cis-eQTL SNP detected with false discovery rate < 0.5 using peripheral blood samples (Westra et al., 2013). In the eQTL/non-eQTL-SNPs set analyzed, all the eQTL and non-eQTL SNPs were selected to be nearly independent of one another ($r^2 \leq 0.1$). In this data set, eQTL SNPs showing stronger associations (i.e., lower *P*-values) with gene expressions were preferentially included, and LD pruning was conducted as in the *P*-value-based pruned sets. Non-eQTL SNPs were randomly selected.

In the DAF-stratified analysis, the allele frequencies of SNPs were determined by the 1000 Genome phase III data (The 1000 Genomes Project Consortium, 2015). For each DAF bin, we used 100,000 SNPs randomly selected from GWAS SNPs regardless of LD. This was because estimates of SP-HMM were unstable due to the small number of SNPs (e.g., a few thousand SNPs) when LD pruned sets were used. Note that, in C4D GWAS, the number of SNPs used in $0.4 < \text{DAF} \leq 0.6$, $0.6 < \text{DAF} \leq 0.8$, and $0.8 < \text{DAF}$ were 94506, 70170, and 49116, respectively, since the SNPs of C4D GWAS was limited (Table S3 in Supplementary Material). The obtained results (i.e., estimates of π and g) using the pruned sets (data not shown) were close to those sampled regardless of LD, and both results had the same trends over DAF bins.

For selecting high quality SNPs and LD information in the above section, HapMap data of Japanese individuals in Tokyo (JPT) and European-ancestry individuals from Utah (CEU) were used for Asian rheumatoid arthritis GWAS data and the other GWAS data, respectively. Similarly, for information of allele frequencies, East Asian and European 1000 Genome Project data were used for Asian rheumatoid arthritis GWAS data and the other GWAS data, respectively.

SOURCE CODE AVAILABILITY

The R code implementing the SP-HMM analysis used in this study is freely available through GitHub (https://github.com/jonishino/SP-HMM).

URLS

HapMap 3, ftp://ftp.ncbi.nlm.nih.gov/hapmap/frequencies/2010-05_phaseIII; 1000 Genome, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502; dbSNP (Build 141), http://www.ncbi.nlm.nih.gov/SNP; eQTL in blood, http://genenetwork.nl/bloodeqtlbrowser/2012-12-21-CisAssociationsProbeLevelFDR0.5.zip; rheumatoid arthritis summary statistics, http://plaza.umin.ac.jp/~yokada/datasource/software.htm; schizophrenia and bipolar disorder summary statistics, www.med.unc.edu/pgc/downloads; coronary artery disease summary statistics, http://www.cardiogramplusc4d.org/.

AUTHOR CONTRIBUTIONS

JN: developed the methods, performed the analyses, and wrote the manuscript. YK: provided essential ideas and interpretations for the study direction and results. DS and TM: contributed to the data acquisition and the analyses. HN: provided the initial version of script for SP-HMM analysis. YK, MK, HO, KB, and TT: improved the manuscript. TT: directed and supervised the study; SM: conceived the study idea, developed the methods, wrote the manuscript, and, directed the study. All authors contributed the final manuscript.

FUNDING

This research was supported by JST CREST Grant Number JPMJCR1412, Japan and a Grant-in-Aid for Scientific Research (16H06299) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

ACKNOWLEDGMENTS

Data on coronary artery disease and myocardial infarction were contributed by CARDIoGRAMplusC4D investigators and were downloaded from www.CARDIOGRAMPLUSC4D.ORG. This study also made use of data generated by Psychiatric Genomic Consortium (PGC). This study has been already published as a preprint, bioRxiv doi: 10.1101/080945 (25 April 2017).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00115/full#supplementary-material

REFERENCES

Agarwala, V., Flannick, J., Sunyaev, S., and Altshuler, D. (2013). Evaluating empirical bounds on complex disease genetic architecture. *Nat. Genet.* 45, 1418–1427. doi: 10.1038/ng.2804

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169, 1177–1186. doi: 10.1016/j.cell.2017.05.038

Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295. doi: 10.1038/ng.3211

Chan, Y., Lim, E. T., Sandholm, N., Wang, S. R., McKnight, A. J., Ripke, S., et al. (2014). An excess of risk-increasing low-frequency variants can be a signal of polygenic inheritance in complex diseases. *Am. J. Hum. Genet.* 94, 437–452. doi: 10.1016/j.ajhg.2014.02.006

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., and Park, J. H. H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* 45, 400–405. doi: 10.1038/ng.2579

Coronary Artery Disease (C4D) Genetics Consortium. (2011). A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat. Genet.* 43, 339–344. doi: 10.1038/ng.782

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390–394. doi: 10.1038/nature10808

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* 29, 51–76. doi: 10.1111/j.1469-1809.1965.tb00500.x

Ferguson, J. P., Cho, J. H., Yang, C., and Zhao, H. (2013). Empirical Bayes correction for the Winner's Curse in Genetic Association Studies. *Genet. Epidemiol.* 37, 60–68. doi: 10.1002/gepi.21683

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47, 1228–1235. doi: 10.1038/ng.3404

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V, Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098. doi: 10.1038/ng.3367

Golan, D., Lander, E. S., and Rosset, S. (2014). Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl. Acad. Sci. U.S.A.* 111, E5272–E5281. doi: 10.1073/pnas.1419064111

Gorlov, I. P., Gorlova, O. Y., and Amos, C. I. (2015). Allelic spectra of risk SNPs are different for environment/lifestyle dependent versus independent diseases. *PLoS Genet.* 11:e1005371. doi: 10.1371/journal.pgen.1005371

Gottesman, I. I., and Shields, J. (1972). A polygenic theory of schizophrenia. *Int. J. Ment. Health* 1, 107–115. doi: 10.1080/00207411.1972.11448568

Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjálmsson, B. J., Xu, H., et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* 95, 535–552. doi: 10.1016/j.ajhg.2014.10.004

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367. doi: 10.1073/pnas.0903103106

International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58. doi: 10.1038/nature09298

Keller, M. C., and Visscher, P. M. (2015). Genetic variation links creativity to psychiatric disorders. *Nat. Neurosci.* 18, 928–929. doi: 10.1038/nn.4047

Lachance, J. (2010). Disease-associated alleles in genome-wide association studies are enriched for derived low frequency alleles relative to HapMap and neutral expectations. *BMC Med. Genomics* 3:57. doi: 10.1186/1755-8794-3-57

Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., et al. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* 44, 247–250. doi: 10.1038/ng.1108

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V, Purcell, S. M., Perlis, R. H., et al. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* 45, 984–994. doi: 10.1038/ng.2711

Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* 88, 294–305. doi: 10.1016/j.ajhg.2011.02.002

Levinson, D. F., Mostafavi, S., Milaneschi, Y., Rivera, M., Ripke, S., Wray, N. R., et al. (2014). Genetic studies of major depressive disorder: why are there no genome-wide association study findings and what can we do about it? *Biol. Psychiatry* 76, 510–512. doi: 10.1016/j.biopsych.2014.07.029

Loh, P. R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* 47, 1385–1392. doi: 10.1038/ng.3431

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

Matsui, S., and Noma, H. (2011a). Estimating effect sizes of differentially expressed genes for power and sample-size assessments in microarray experiments. *Biometrics* 67, 1225–1235. doi: 10.1111/j.1541-0420.2011.01618.x.

Matsui, S., and Noma, H. (2011b). Estimation and selection in high-dimensional genomic studies for developing molecular diagnostics. *Biostatistics* 12, 223–233. doi: 10.1093/biostatistics/kxq057

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Eileen Dolan, M., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6:e1000888. doi: 10.1371/journal.pgen.1000888

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506, 376–381. doi: 10.1038/nature12873

Palla, L., and Dudbridge, F. (2015). A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *Am. J. Hum. Genet.* 97, 250–259. doi: 10.1016/j.ajhg.2015.06.005

Park, J. H., Gail, M. H., Weinberg, C. R., Carroll, R. J., Chung, C. C., Wang, Z., et al. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl. Acad. Sci. U.S.A.* 108, 18026–18031. doi: 10.1073/pnas.1114759108

Park, J. H. H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., et al. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* 42, 570–575. doi: 10.1038/ng.610

Power, R. A, Steinberg, S., Bjornsdottir, G., Rietveld, C. A., Abdellaoui, A., Nivard, M. M., et al. (2015). Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat. Neurosci.* 18, 953–955. doi: 10.1038/nn.4040

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 10, 8192–8192. doi: 10.1038/nature08185.

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K.-H., Holmans, P. A., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427. doi: 10.1038/nature13595

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kahler, A. K., Akterin, S., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* 45, 1150–1159. doi: 10.1038/ng.2742

Sawyer, S. A., and Hartl, D. L. (1992). Population genetics of polymorphism and divergence. *Genetics* 132, 1161–1176.

Schunkert, H., Konig, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* 43, 333–338. doi: 10.1038/ng.784

Shen, W., and Louis, T. A. (1999). Empirical Bayes estimation via the smoothing by roughening approach. *J. Comput. Graph. Stat.* 8, 800–823.

Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.* 99, 139–153. doi: 10.1016/j.ajhg.2016.05.013

Sklar, P., Ripke, S., Scott, L. J., Andreassen, O., Cichon, S., Craddock, N., et al. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* 43, 977–983. doi: 10.1038/ng.943

So, H. C., Gui, A. H. S., Cherny, S. S., and Sham, P. C. (2011a). Evaluating the heritability explained by known susceptibility variants: A survey of ten complex diseases. *Genet. Epidemiol.* 35, 310–317. doi: 10.1002/gepi.20579

So, H. C., Li, M., and Sham, P. C. (2011b). Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. *Genet. Epidemiol.* 35, 447–456. doi: 10.1002/gepi.20593

Stahl, E. A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B. F., et al. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* 44, 483–489. doi: 10.1038/ng.2232

Stephens, M., and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* 10, 681–690. doi: 10.1038/nrg2615

Stringer, S., Kahn, R. S., de Witte, L. D., Ophoff, R. A., and Derks, E. M. (2014). Genetic liability for schizophrenia predicts risk of immune disorders. *Schizophr. Res.* 159, 347–352. doi: 10.1016/j.schres.2014.09.004

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006. doi: 10.1093/nar/gkt1229

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* 45, 1238–1243. doi: 10.1038/ng.2756

Wray, N., Goddard, M., and Visscher, P. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* 17, 1520–1528. doi: 10.1101/gr.6665407

Wray, N. R., Pergadia, M. L., Blackwood, D. H. R., Penninx, B. W. J. H., Gordon, S. D., Nyholt, D. R., et al. (2012). Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Mol. Psychiatry* 17, 36–48. doi: 10.1038/mp.2010.109

Yang, J., Wray, N. R., and Visscher, P. M. (2010). Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genet. Epidemiol.* 34, 254–257. doi: 10.1002/gepi.20456.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* 48, 481–487. doi: 10.1038/ng.3538

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Nishino, Kochi, Shigemizu, Kato, Ikari, Ochi, Noma, Matsui, Morizono, Boroevich, Tsunoda and Matsui. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.