bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Empirical Bayes estimation of semi-parametric hierarchical mixture models for unbiased characterization of polygenic disease architectures

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

[1]Department of Biostatistics, Nagoya University Graduate School of Medicine, Nagoya, Aichi, Japan; [2]Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan; [3]Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan; [4]Department of Bioinformatics, National Cancer Center Research Institute, Tokyo, Japan; [5]Institute of Rheumatology, Tokyo Women's Medical University, Tokyo, Japan; [6]Department of Medicine and Molecular Science, Division of Frontier Medical Science, Programs for Biomedical Research, Graduate School of Biomedical Sciences, Hiroshima University, Hiroshima, Japan; [7]Department of Data Science, The Institute of Statistical Mathematics, Tachikawa, Tokyo, Japan; [8]Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan; [9]CREST, JST, Japan

Send all correspondence to:

Shigeyuki Matsui, Ph.D.

Department of Biostatistics, Nagoya University Graduate School of Medicine,

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

Tel.: +81-52-744-2487, Fax.: +81-52-744-2488

E-mail: smatsui@med.nagoya-u.ac.jp

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Empirical Bayes estimation of semi-parametric hierarchical mixture models for unbiased characterization of polygenic disease architectures

Abstract

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 architecture of many diseases has not been accurately assessed 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 (~30% risk variants of independent genome-wide SNPs, most within odds ratio (OR)=1.03), whereas rheumatoid arthritis was less polygenic (2.5~5.0% risk variants, significant portion reaching OR=1.05~1.1). 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 mutation. 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.

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome-wide association studies (GWAS) have identified numerous susceptibility variants for complex diseases[1]. 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[2]. Many researches have 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[3–11].

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[7,8,12–16]. The estimation of these components also contributes to accurate power and sample size calculations of GWAS[7,12,13,17–19] and estimation of the predictive capability of disease risks[12,15,16].

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 to unbiasedly evaluate various polygenic architectures using the entire 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[8]. Its application, however, has been limited to few studies[7,8] because of technical complexity and excess computational burden with many simulations. On the other hand, some authors estimated the effect-size distribution based on a power evaluation for SNPs reaching genome-wide significance[13–15]. 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 the existing methods, we propose an

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

empirical Bayes estimation of semi-parametric hierarchical mixture models (SP-HMMs)[20,21] of GWAS summary statistics on effect sizes, such as estimated log-odds ratios to associate genotypes with disease susceptibility. Mixture modelling 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 non-null distribution, semi-parametric hierarchical modelling 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[22], we can accurately 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[3] and adjustment for the winner's curse[23].

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 accurately estimated, 2) it requires only summary data from GWAS or meta-analysis of 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.

Throughout this paper, we fit the SP-HMM to summary data from meta-/mega-analyses for various diseases such as rheumatoid arthritis[24], schizophrenia[25], bipolar disorder[26], and coronary artery disease[27,28], to estimate the respective polygenic architectures and compare them across diseases. We also assess the liability-scale variance explained by SNPs based on this estimation. In order to obtain further insight into the underlying polygenic architectures, our approach can easily incorporate important functional

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

categories, such as expression quantitative trait loci (eQTL), coding, non-synonymous, promoter, 5' or 3' UTR, enhancer, and DNase I hypersensitivity sites[29–32]. We focus on eQTLs, as gene expression levels have been increasingly recognized as notable endophenotypes or important mediators between genetic variations and disease phenotypes[29,30,33,34]. Lastly, we also incorporate derived allele frequency (DAF), rather than minor allele frequency (MAF)[14,35,36]. 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 by negative selection. Thus, our DAF-based analysis facilitates interpretation from the perspective of population genetics[37], possibly contributing to further understanding of the genetic etiology for complex diseases.

RESULTS

We first confirmed the adequacy of our estimation method in unbiasedly estimating the proportion of disease-associated SNPs, π, and the effect-size distribution, g, in simulation experiments (see Supplementary Note, Supplementary Tables 1-5 and Supplementary Figs. 1-4). The non-parametric estimation for g could flexibly capture various forms of the underlying effect-size distributions.

For application to real GWAS datasets, we used publicly available summary statistics from large meta-/mega-GWAS for the four complex diseases (see Supplementary Tables 6 and 7 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

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To estimate the proportion of disease-associated SNPs, π, and the effect-size distribution, g, based on independent SNPs, we used two pruned SNP sets: P-value-based and random-pruned sets. 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 Online Methods for details). The random-pruned method sampled SNPs randomly. In both methods, pruned SNPs with linkage disequilibrium (LD, $r^2 \leq 0.1$) were selected. Of note, one causal variant would not be redundantly tagged by SNPs in the pruned SNP sets, whereas not all causal variants would be well tagged by SNPs even in the P-value-based sets. Thus, the estimates $\hat{\pi}$ × (the number of SNPs in the SNP sets) using the pruned sets would give conservative estimates of the number of causal variants.

We fit the SP-HMM to the P-value-based pruned SNP sets in each GWAS (Table 1; Fig. 1). For rheumatoid arthritis, π was estimated as 2.5% for Asian and 5.0% for European populations, which were lower than the other diseases. The estimates of π were larger for two psychiatric diseases: 31.7% for schizophrenia and 20.5% for bipolar disorder. For coronary artery disease, using CARDIoGRAM and C4D data, π was estimated to be 8.0% and 10.7%, respectively.

With regard to the estimation of g, rheumatoid arthritis was shown to have a distribution with larger effects, spanning to $|\beta|$ = 0.05 (odds ratio of 0.95 or 1.05) or larger, when compared with the other diseases (Fig. 1). Interestingly, we also observed that a group of SNPs with very large effects at $|\beta|$ > 0.05 formed clear peaks. 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 was shown to have the narrowest distribution with very small effects.

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Coronary artery disease was also estimated to have a distribution with relatively small effects, next to schizophrenia.

The estimates of π for the random-pruned SNP sets were similar to those for the P-value-based SNP sets for each GWAS (Supplementary Table 8). 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 (Supplementary Fig. 5).

Liability-scale variance explained by SNPs

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, explained by SNPs (i.e., SNP heritability). 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 20.0% 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.[8]), 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%[8]. The estimates of V in schizophrenia and bipolar disorder were around 40% and 50%, respectively, which were higher but almost within the range of previously reported estimates of 23-43% and 37-47%, respectively, for these diseases[5,7,9,10]. For cardiovascular disease, the estimates of V from the CARDIoGRAM and C4D data were 20.7% and 22.0%, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Supplementary Table 8).

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[38], and the remaining 'non-eQTL'-SNPs, separately (Fig. 2). 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 Online 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 (Fig. 2). 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 Fig. 2 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 (Supplementary Table 9). For rheumatoid arthritis, as expected from Fig. 2, the per-SNP variance for the eQTL-SNPs was much larger than for the non-eQTL-SNPs. Interestingly, although 'eQTL' was defined using European samples[38], the enrichment of per-SNP variance (10.7-fold) in the eQTL-SNPs in the Asian population was larger than the 5.7-fold enrichment seen in the European population.

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Estimation across derived allele frequencies

The effect size estimation of GWAS data stratified with the derived allele frequency (DAF) could provide another perspective on polygenic architecture, which facilitates assessment based on population genetics (see Discussion). We classified all SNPs into ten 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 (Figs. 3a, b). We observed sharp peaks at positive effects, i.e., $\beta > 0$, for lower DAF bins, especially for DAF $\leq$ 0.1, and at negative effects for higher DAF bins, especially for DAF > 0.9. This indicates that low-frequency-derived and high-frequency-derived alleles are prone to act as risk and protective variants for disease occurrence, respectively. Further, the proportions of disease-associated SNPs were clearly higher at DAF $\leq$ 0.1 and DAF > 0.9 (Supplementary Fig. 6). 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 < 0.1 and DAF > 0.9, respectively, whereas, for bipolar disorder, negative and positive effects were over-represented at DAF $\leq$ 0.1 and DAF > 0.9, respectively.

DISCUSSION

We have developed a simple and fast method for unbiasedly estimating the proportion of disease-associated variants and the effect-size distribution based on the empirical Bayes estimation of SP-HMM. The proposed method can effectively distinguish various polygenic architectures, including the degree of polygenicity, across diseases, and can also provide

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

various perspectives of the polygenic architecture based on important variant categories such as DAF and eQTL.

Schizophrenia, which has been suspected to be highly polygenic[39,40], was estimated to have ~ 30% disease-associated variants with very small effects (most within $|\beta|$ = 0.03) of independent SNPs in the genome (Table 1, Fig. 1, Supplementary Table 8, and Supplementary Fig. 5). This suggests at least ~ 30,000 causal variants exist in the genome, which does not contradict a recent study that estimated at least ~ 20,000 causal variants by a simulation-based method[41]. In contrast, rheumatoid arthritis was found to be less polygenic. Our estimates of π, 2.5% for Asians and 5.0% for Europeans, were generally consistent with previous estimates of 2.7%[8] or 5.4%[10] for Europeans, and the estimate for g ranged to $|\beta|$ = 0.05 or even 0.1 (Table 1 and Fig. 1). In fact, the effect sizes of validated variants for rheumatoid arthritis were generally larger than those for schizophrenia[24,25]. Our estimate of g means that the effect sizes of variants that can be detected would also be relatively larger among complex diseases.

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 (Supplementary Table 9), similar to the enrichment on per-SNP variance by blood-specific DNaseI hypersensitivity sites (DHS)[33], which were also strongly associated with expression variation[42]. 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 'eQTL' was defined using European samples[38], 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 Fig. 5 in Okada et al.[24]). This might be explained by

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 $|\beta| > 0.05$ (Fig. 2) in rheumatoid arthritis.

The SP-HMM can also provide posterior effect-size estimates of individual SNPs based on the estimated genetic architecture, $\hat{\pi}$ and $\hat{g}$.[20,43] 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, $|\beta| > 0.05$ (Supplementary Data Set 1). As this list includes eQTLs such as RNASET2 and ADO, which have not been previously linked to rheumatoid arthritis[24], 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 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 that genetic overlap between immune diseases and schizophrenia has been previously reported[44]. However, it should be noted that precise estimation of the eQTL effects in these diseases needs additional eQTL data covering all the tissues and cells related to the diseases.

Using DAF-stratified analysis for rheumatoid arthritis, we estimated more risk/protective derived alleles in low/high DAF (Fig. 3) and larger π in low or high DAF (Supplementary Fig. 6). Simple models based on a theory of population genetics for DAF[45] (see Supplementary Fig. 7) could help interpret results from the DAF analysis, and thus provide another perspective on the difference among diseases (see Supplementary Note for details). Among such models, the 'deleterious-risk and advantageous-protective mutation' model with weak selection was best fitted for rheumatoid arthritis (Supplementary Fig. 8). Because most

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of the risk genes for rheumatoid arthritis are implicated in immune system regulation[24], these low- and high-derived alleles would tend to skew individual's immune function towards 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[6,40], we observed opposite tendencies in the genetic architecture for these diseases: risk (protective) and protective (risk) derived alleles were over-represented, especially at DAF $\leq$ 0.1 and DAF $>$ 0.9 for schizophrenia (bipolar disorder) (Fig. 3). 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. [35]). 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 (Supplementary Fig. 7). Recently, genetic correlations between creativity and both schizophrenia and bipolar disorder were reported, but they were much stronger for bipolar disorder[46,47]. Taken together with our estimation, these results might provide a clue for resolving the shared and specific genetic etiologies between the two genetically related diseases.

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[7,13] and estimation of predictive capability of disease risk[15]. The SP-HMM can also be extended to multi-dimensional settings, e.g., for quantification of sex or ethnic differences in

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genetic architecture for a disease, or (antagonistic) pleiotropic genetic architecture in multiple diseases. This kind of multi-dimensional analysis is 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 reported in future reports.

REFERENCE

1. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–D1006 (2014).

2. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

3. So, H. C., Li, M. & Sham, P. C. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. *Genet. Epidemiol.* **35,** 447–456 (2011).

4. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88,** 294–305 (2011).

5. Lee, S. H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44,** 247–250 (2012).

6. Lee, S. H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45,** 984–994 (2013).

7. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45,** 1150–1159 (2013).

8. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44,** 483–489 (2012).

9. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: Inferring the contribution of common variants. *Proc. Natl. Acad. Sci.* **111,** E5272–E5281 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

10. Palla, L. & Dudbridge, F. 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 (2015).

11. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).

12. Wray, N., Goddard, M. & Visscher, P. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17,** 1520–1528 (2007).

13. Park, J.-H. H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42,** 570–575 (2010).

14. Park, J.-H. *et al.* 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 (2011).

15. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45,** 400–405 (2013).

16. Agarwala, V., Flannick, J., Sunyaev, S. & Altshuler, D. Evaluating empirical bounds on complex disease genetic architecture. *Nat. Genet.* **45,** 1418–1427 (2013).

17. Yang, J., Wray, N. R. & Visscher, P. M. Comparing apples and oranges: Equating the power of case-control and quantitative trait association studies. *Genet. Epidemiol.* **34,** 254–257 (2010).

18. Wray, N. R. *et al.* Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Mol. Psychiatry* **17,** 36–48 (2012).

19. Levinson, D. F. *et al.* 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 (2014).

20. Matsui, S. & Noma, H. Estimation and selection in high-dimensional genomic studies for developing molecular diagnostics. *Biostatistics* **12,** 223–233 (2011).

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

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

23. Ferguson, J. P., Cho, J. H., Yang, C. & Zhao, H. Empirical Bayes Correction for the Winner's Curse in Genetic Association Studies. *Genet. Epidemiol.* **37,** 60–68 (2013).

24. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506,** 376–381 (2014).

25. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).

26. Sklar, P. *et al.* Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* **43,** 977–983 (2011).

27. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* **43,** 333–338 (2011).

28. Consortium., T. C. A. D. (C4D) G. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat. Genet.* **43,** 339–344 (2011).

29. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genet.* **6,** e1000888 (2010).

30. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47,** 1091–1098 (2015).

31. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci.* **106,** 9362–9367 (2009).

32. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).

33. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95,** 535–552 (2014).

34. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48,** 481–487 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

35. Chan, Y. *et al.* 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 (2014).

36. Gorlov, I. P., Gorlova, O. Y. & Amos, C. I. Allelic spectra of risk SNPs are different for environment/lifestyle dependent versus independent diseases. *PLoS Genet.* **11,** e1005371 (2015).

37. Lachance, J. 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 (2010).

38. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).

39. Gottesman, I. I. & Shields, J. A polygenic theory of schizophrenia. *Int. J. Ment. Health* **1,** 107–115 (1972).

40. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **10,** 8192–8192 (2009).

41. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47,** 284–290 (2015).

42. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (2012).

43. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10,** 681–690 (2009).

44. Stringer, S., Kahn, R. S., de Witte, L. D., Ophoff, R. a. & Derks, E. M. Genetic liability for schizophrenia predicts risk of immune disorders. *Schizophr. Res.* **159,** 347–352 (2014).

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

46. Power, R. a *et al.* Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat. Neurosci.* **18,** 953–955 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

47. Keller, M. C. & Visscher, P. M. Genetic variation links creativity to psychiatric disorders. *Nat. Neurosci.* **18,** 928–929 (2015).

48. So, H. C., Gui, A. H. S., Cherny, S. S. & Sham, P. C. Evaluating the heritability explained by known susceptibility variants: A survey of ten complex diseases. *Genet. Epidemiol.* **35,** 310–317 (2011).

49. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29,** 51–76 (1965).

50. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

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

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/.

ONLINE 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'

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 Y_j's, we assumed a mixture structure with two components, null and non-null SNPs, in terms of association with disease susceptibility. To be specific,

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

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 $y_j \sim f_{0j}(y_j) = N\left(0, \hat{V}_{\hat{\beta}_j}\right)$ 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: $y_j|\beta_j \sim f_{1j}(y_j|\beta_j) = N\left(\beta_j, \hat{V}_{\hat{\beta}_j}\right)$ 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)[20,21]. The assumption that y_j's are 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[22], to incorporate the non-parametric prior distribution g[20,21]. 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

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

complex diseases. We set the number grid points as 120, such that $\boldsymbol{b}$ = (−0.300, −0.295, ..., −0.005, 0.005, ..., 0.295, 0.300). The recursion of the EM algorithm was stopped when the relative changes in the estimate of π were small and the first two moments of g became stable, following the guideline by Shen and Louis[22]. 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

The log odds ratio, β_j, together with the allele frequency and the disease prevalence, were transformed to the variance explained by the j-th SNP, denoted as v_j, in the liability threshold model[48]. 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 the Hardy-Weinberg equilibrium in the population, the genotype frequencies are given by $h_{j,0} = (1 - p_j)^2$, $h_{j,1} = 2p_j(1 - p_j)$, 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

$$v_j{}^* = \sum_{x_j=0}^{2} h_{j,x_j} \, (\mu_{j,x_j} - \mu_{all})^2.$$

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

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

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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[49,4]

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 Genome phase III[50] and the same prevalences as previously assumed in estimating SNP heritability were used[8,9]. 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 (see URLs). The characteristics of individual GWASs are shown in Supplementary Tables 1 and 2. To restrict analysis to well-imputed, high-quality variants, we used only SNPs that existed on the HapMap 3 reference panel[51] and the 1000 Genome phase III data[50]. For rheumatoid arthritis, the MHC region (chromosome 6, 25 – 35 Mb) was removed. The derived/ancestral states of alleles were determined by using the 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 Fig. 1). 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

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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). 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 'eQTL SNP' as cis-eQTL SNPs 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 DAF-stratified analysis, the allele frequencies of SNPs were determined by the 1000 Genome phase III data[50]. 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 small number of SNPs (e.g., a few thousand SNPs) when LD pruned sets were used. 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)

bioRxiv preprint doi: https://doi.org/10.1101/080945; this version posted October 21, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Acknowledgments

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

Author Contributions

J.N. developed the methods, performed the analyses, and wrote the manuscript. Y.K. provided essential ideas and interpretations for the study direction and results. D.S. and T.M. contributed to the data acquisition and the analyses. H.M. provided the initial version of script for SP-HMM analysis. Y.K., M.K., H.O., K. A. B., and T.T. improved the manuscript. T.T. directed and supervised the study. S.M. conceived the study idea, developed the methods, wrote the manuscript, and, directed the study. All authors contributed the final manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

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)	2.5 (0.5)	14.0 (1.8)
Rheumatoid arthritis (European)	5.0 (0.6)	20.0 (1.4)
Coronary artery disease (CARDIoGRAM)	8.0 (0.7)	20.7 (1.2)
Coronary artery disease (C4D)	10.7 (0.9)	22.0 (1.2)
Schizophrenia	31.7 (0.4)	40.2 (0.8)
Bipolar disorder	20.5 (0.6)	50.0 (1.7)

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.

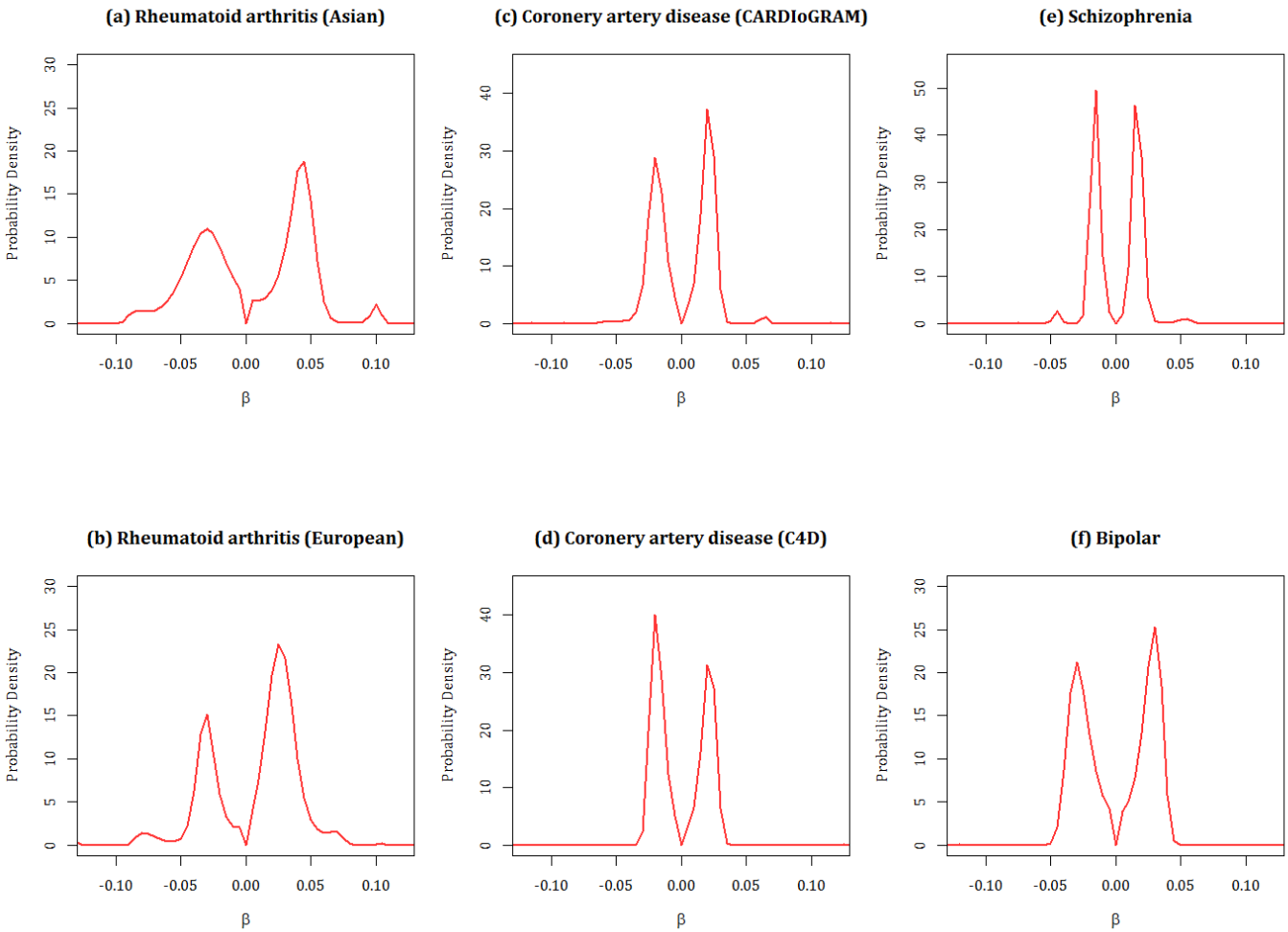

Figure 1 Estimated effect-size distributions for disease-associated SNPs, $\hat{g}$. The P-value-based pruned SNP sets are used.

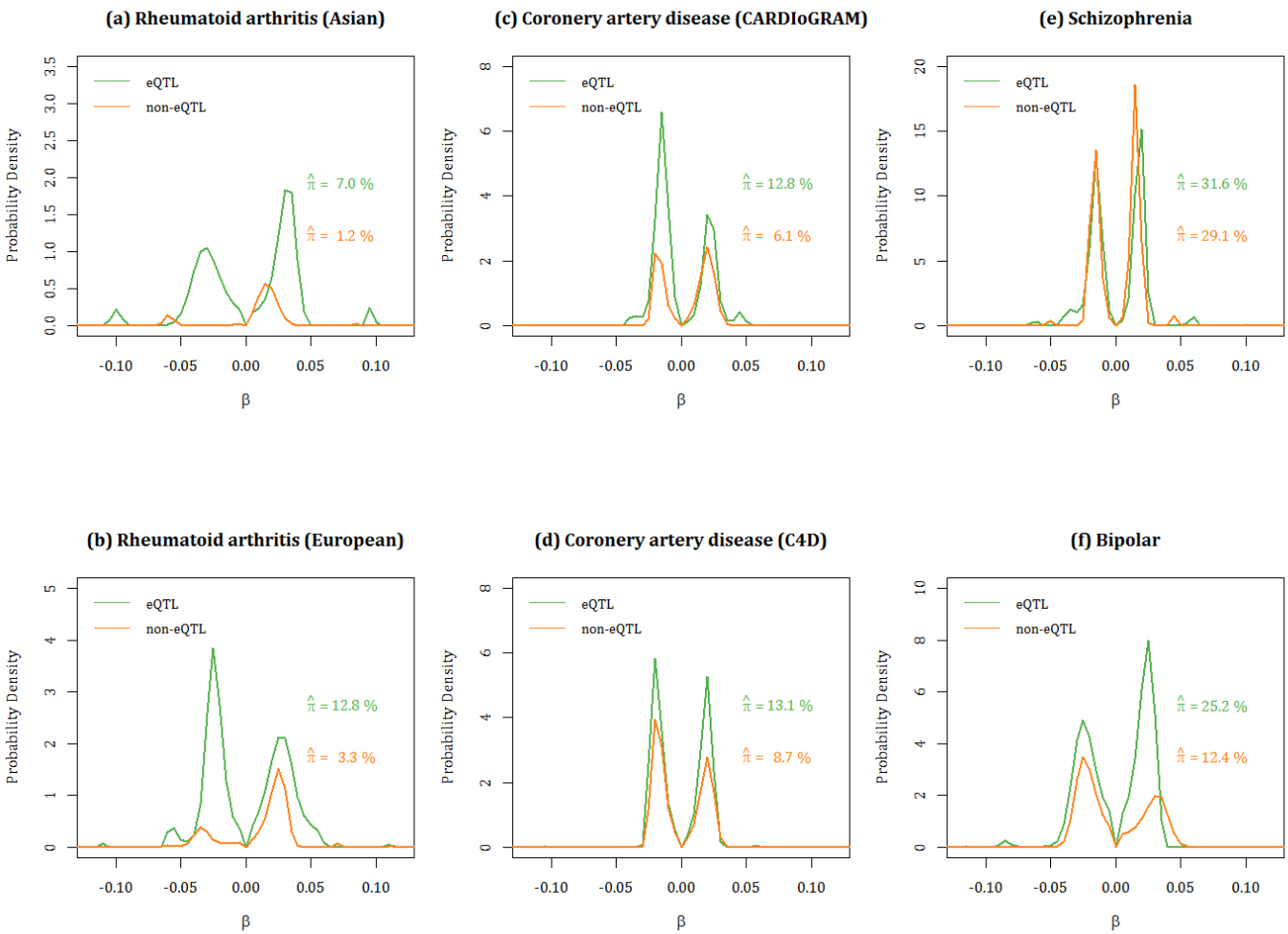

Figure 2 Estimated effect size distributions for eQTL-SNPs and non-eQTL-SNPs, $\hat{\pi} \times \hat{g}$. Green and orange lines show the results for the eQTL-SNP and non-eQTL-SNP sets, respectively. Estimated proportions of disease-associated SNPs, $\hat{\pi}$, correspond to the areas under the curves.

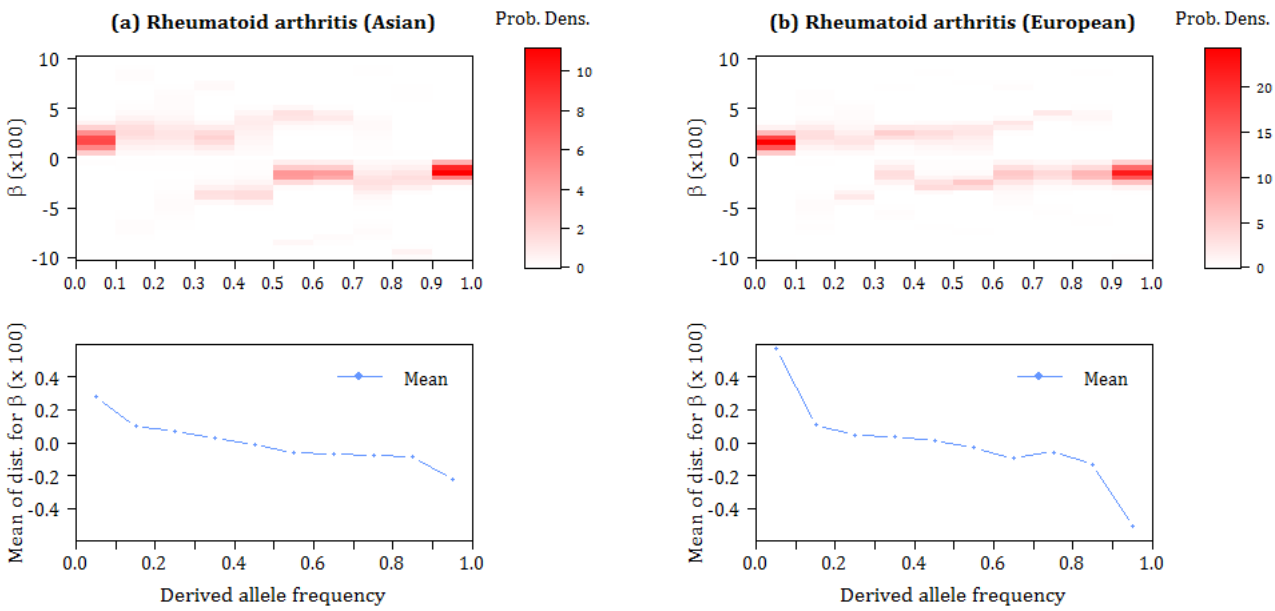

Figure 3 Estimated effect size distributions , $\hat{\pi} \times \hat{g}$, by derived allele frequency (DAF) bins. The upper panels (heatmap colors) for each GWAS results show $\hat{\pi} \times \hat{g}$. The lower panels show means of $\hat{\pi} \times \hat{g}$.

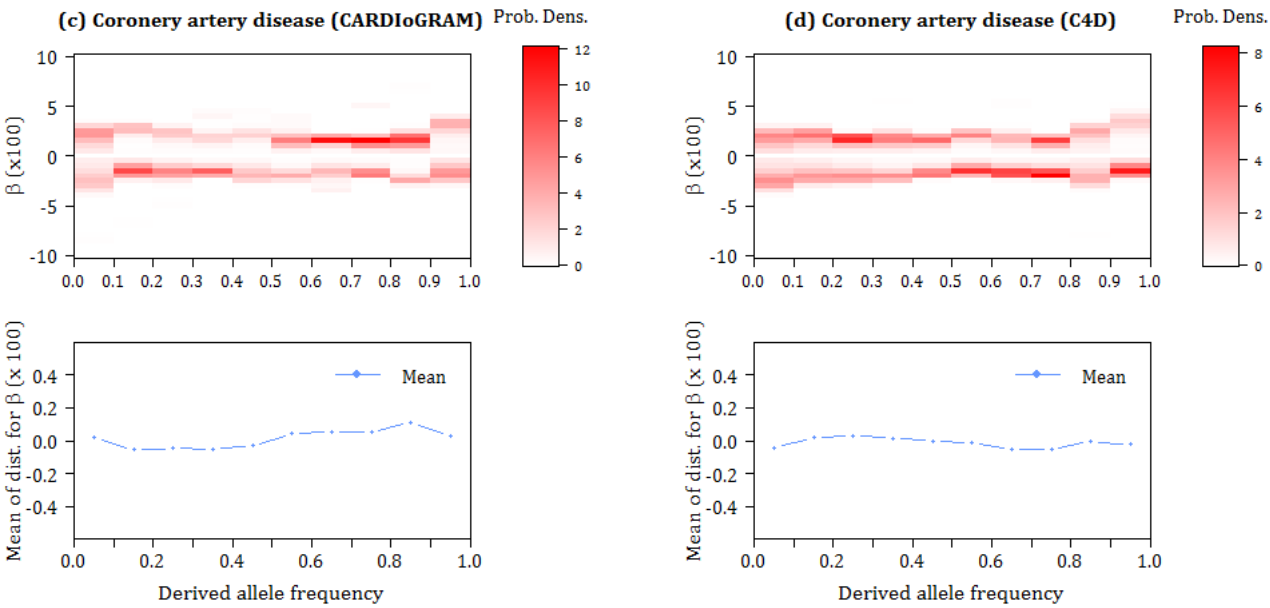

Figure 3 Continued.

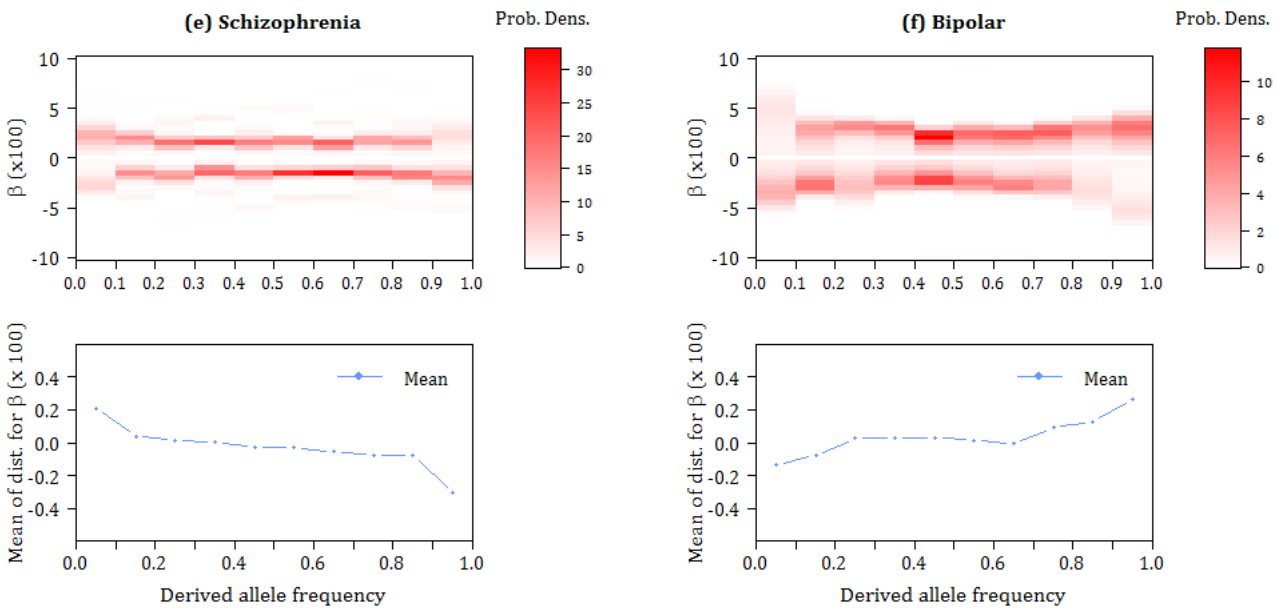

Figure 3 Continued.