European Journal of Human Genetics (2018) 26:1038–1048
https://doi.org/10.1038/s41431-018-0125-3

ARTICLE

Re-assessment of multiple testing strategies for more efficient genome-wide association studies

Takahiro Otani[1] · Hisashi Noma[2] · Jo Nishino[3] · Shigeyuki Matsui[3]

Received: 19 January 2017 / Revised: 19 December 2017 / Accepted: 13 February 2018 / Published online: 9 March 2018

Abstract

Although enormous costs have been dedicated to discovering relevant disease-related genetic variants, especially in genome-wide association studies (GWASs), only a small fraction of estimated heritability can be explained by these results. This is the so-called missing heritability problem. The conventional use of overly conservative multiple testing strategies based on controlling the familywise error rate (FWER), in particular with a genome-wide significance threshold of $P < 5 \times 10^{-8}$, is one of the most important issues from a statistical perspective. To help resolve this problem, we performed comprehensive re-assessments of currently available strategies using recently published, extremely large-scale GWAS data sets of rheumatoid arthritis and schizophrenia (>50,000 subjects). The estimates of statistical power averaged for all disease-related genetic variants of the standard FWER-based strategy were only 0.09% for the rheumatoid arthritis data and 0.04% for the schizophrenia data. To design more efficient strategies, we also conducted an extensive comparison of multiple testing strategies by applying false discovery rate (FDR)-controlling procedures to these data sets and simulations, and found that the FDR-based procedures achieved higher power than the FWER-based strategy, even at a strict FDR level (e.g., FDR = 1%). We also discuss a useful alternative measure, namely "partial power," which is an averaged power for detecting the clinically and biologically meaningful genetic factors with the largest effects. Simulation results suggest that the FDR-based procedures can achieve sufficient partial power (>80%) for detecting these factors (odds ratios of >1.05) with 80,000 subjects, and thus this may be a useful measure for defining realistic objectives of future GWASs.

Introduction

Due to advances in high-throughput technology in medicine and molecular biology, enormous costs and resources have been dedicated to discovering relevant disease-related genetic variants, especially in genome-wide association studies (GWASs) conducted over the past decade [1]. However, almost none of these studies could identify more than several dozen variants so as to reach the genome-wide significance level [2], and in many diseases only a small fraction of the estimated heritability could be explained by these variants, the so-called missing heritability problem [3]. To overcome this issue, several "mega"-analyses have aggregated large-scale GWAS data sets to gain statistical efficiency (>50,000 subjects), but they were able to detect only small numbers of additional significant variants [4–6].

Although these problems may be due to various complex factors, statistical analysis strategies are among the most important issues. In particular, multiple testing strategies that control the familywise error rate (FWER) have been typically applied in GWAS to adjust the multiplicity of millions of statistical tests [2, 7, 8]. In principle, the FWER criterion strictly controls the probability of having at least one false positive in millions of tests, and geneticists should generally recognize its inappropriateness regarding the primary purposes of GWAS, i.e., screening candidates with relevant variants for further investigations. Other criteria such as the false discovery rate (FDR) [9], the expected

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41431-018-0125-3) contains supplementary material, which is available to authorized users.

✉ Hisashi Noma
noma@ism.ac.jp

[1] Risk Analysis Research Center, The Institute of Statistical Mathematics, Tachikawa, Tokyo 190-8562, Japan

[2] Department of Data Science, The Institute of Statistical Mathematics, Tachikawa, Tokyo 190-8562, Japan

[3] Department of Biostatistics, Nagoya University Graduate School of Medicine, Nagoya, Aichi 466-8550, Japan

proportion of false positives to total significant results, have been proposed and well discussed [10–12], but almost all GWASs have still used FWER-controlling strategies (in particular, the genome-wide significance threshold [2] of $P < 5 \times 10^{-8}$).

In this article, we present a comprehensive re-assessment of multiple testing strategies to gain insight into overcoming the missing heritability problem from a statistical perspective. In particular, we conducted extensive comparisons of currently available strategies using large-scale GWAS data sets from recently published mega-analyses of rheumatoid arthritis [4] and schizophrenia [5]. Surprisingly, our estimates of statistical power indicated that even for the largest data sets that had extensively aggregated the currently available GWAS resources, the standard FWER-controlling strategy (using the genome-wide significance threshold) was able to discover less than 0.1% of the total relevant genetic factors (see "Statistical power of the multiple testing strategies for the large-scale GWAS data sets"). This suggests the serious problem that most disease-related genetic factors will not be identified if current GWAS practices are continued, and much greater costs and resources will be dedicated to future studies aggregating >100,000 subjects. Specifically, we do not know for certain how efficient the current multiple testing practices are, or how accurately these methods can identify disease-related genetic factors under the realistic conditions of current GWASs. If the conventional frameworks do not work efficiently, it is time that we consider changing the statistical analysis practices used in GWASs.

To address these problems, we also conducted large-scale simulation studies to evaluate the efficiency of multiple testing strategies, in particular estimating how large statistical power will be if the same practices are continued and the cumulative sizes of GWAS data sets become much larger (~80,000 subjects). Although several simulation studies comparing the effectiveness of multiple testing procedures for GWASs have already been reported [13, 14], these studies were small in scale, compared only a few methods under limited conditions, and did not consider highly polygenic diseases in which a number of disease-related single nuclear polymorphisms (SNPs; >1% of all SNPs) with modest effects (odds ratios (ORs) of <1.05) contribute to disease risk. By contrast, we considered more realistically scaled GWASs, with a total of several tens of thousands of subjects with a million genotyped SNPs, and focused on association studies of highly polygenic diseases that are the main target of recent and future GWASs. Under these conditions, we comprehensively compared the expected statistical power of currently available multiple testing strategies that have widely been used in genomics.

In addition, the estimated statistical power of large-scale mega-analyses of GWASs (less than 0.1%; see "Statistical power of the multiple testing strategies for the large-scale GWAS data sets") also suggests another relevant question, namely, whether the objective of "discovering all (or a certain large fraction) of disease-related genetic factors" is achievable, at least under the realistic current practice conditions of GWASs. This issue indicates that the conventional statistical power index might not be adequate for designing and assessing a practical strategy for GWASs. In this article, we also discuss alternative useful measures [15] for using extensive numerical evaluations to explore efficient GWAS strategies.

Materials and methods

GWAS data sets

To evaluate the efficiency of multiple testing strategies in current large-scale GWAS data sets, we used recently published GWAS data sets of rheumatoid arthritis [4] and schizophrenia [5]. These data sets were obtained from the largest-scale meta-analysis and "mega"-analysis (joint analyses of pooled individual-level data) of these diseases, respectively, with a total of >50,000 subjects and combining several GWAS data sets. For the rheumatoid arthritis data set, we used a data set from a GWAS meta-analysis of four studies of Asian populations. Also, the GWAS of schizophrenia was conducted in individuals of European ancestry. The potential population substructure was corrected by principal components analysis in each study [4, 5]. The numbers of SNPs used in the evaluation are 6,318,961 for the rheumatoid arthritis data set and 1,252,901 for the schizophrenia data set. For details, see Section A in the Supplementary Notes.

Multiple testing strategies

Due to the overly conservative principles of the standard FWER-controlling strategy, which strictly controls the probability of having at least one false positive, approaches based on the FDR have been widely discussed and practically applied in other fields (e.g., gene expression microarray analyses [16]) as an alternative, but comprehensive assessments of these strategies in the context of current GWASs have not been well investigated. To evaluate their applicability, we conducted extensive comparisons of currently available strategies by applying them to the current GWAS data sets and simulation studies. We assessed not only the performance of the FWER-controlling strategy, but also those of five FDR-controlling procedures chosen as representatives of various FDR frameworks involving two frequentist FDR algorithms: the Benjamini–Hochberg procedure [9] and Storey's procedure [11, 17]; the optimal discovery procedure [18, 19]; "Locfdr [20, 21]," an empirical Bayes approach; and a recently developed Wakefield's Bayesian method for GWASs [22–24]. We implemented these procedures in the R Statistical Programming Language (available at https://www.r-project.org/) using widely distributed optional packages (see Section B in the Supplementary Notes for details).

Power diagnostics

To assess the efficiency of the various testing strategies, we considered two types of statistical power measures, considering especially the definition of achievable objectives under realistic GWAS situations. First, we considered the conventional averaged power [25, 26], referred to as "overall power," defined as the expected proportion of true positives among true alternatives for a given significance criterion. This power can be used to estimate the potential discoveries of future GWASs as conducted in recent studies [27, 28], and provides a usual measure in that it reflects the intent to detect all disease-related genetic factors regardless of their effect sizes. However, the estimated overall power of the standard FWER-controlling strategy for large-scale mega-analyses of GWASs was extremely low (less than 0.1%; see "Statistical power of the multiple testing strategies for the large-scale GWAS data sets"), and it would generally be difficult to detect all relevant factors for practically acceptable sample sizes (even if they exceed 100,000). These facts suggest that the overall power might not be adequate for defining realistic objectives of GWASs.

A practical compromise for addressing this issue would be to consider that detecting genetic factors with the largest effect sizes is more important than detecting others with smaller effect sizes. Geneticists and clinicians have particular interest in such factors in order to elucidate disease biology or develop effective risk prediction algorithms. Therefore, we could also consider an alternative measure that focuses on disease-related genetic factors with large effect sizes. Matsui and Noma [15] proposed another useful measure, "partial power," that is defined as the averaged power of all genetic factors with effect sizes larger than a pre-specified threshold. Naturally, the statistical power for individual variants differs according to these variants' effect sizes, and sufficient averaged power could be achieved by focusing in particular on the largest subsets of the disease-related variants under practically feasible sample sizes (see "Simulation studies for evaluating statistical power for future GWASs"). In addition, this is a rather realistic objective that geneticists and clinicians have already aimed toward, and would provide a new perspective for designing GWAS strategies. In practice, the partial power can be used in the same way as the overall power for designing studies and for assessing efficiency [27]. We evaluated these power indices under specified levels of FDR. To estimate these measures under general GWAS settings, we can obtain model-based estimates under an empirical Bayes framework based on a semi-parametric hierarchical mixture model [15, 29] (see Sections C and D in the Supplementary Notes for technical details).

Simulation studies for evaluating false positive rates and statistical power

Based on the preceding discussion, we evaluated the efficiency of multiple testing strategies in detail by conducting comprehensive simulation studies that imitated large-scale mega-analyses of GWASs of polygenic diseases. In particular, we investigated the expected sizes of statistical power if current practices are further continued and the cumulative sizes of GWAS data sets become much larger. PLINK [30] (available at http://pngu.mgh.harvard.edu/~purcell/plink/) was used to generate a large-scale GWAS data set for case–control studies. The number of SNPs was set at 1,000,000 throughout the simulations. We considered equal numbers of cases and controls, with sample sizes of $N =$ 10,000, 20,000, …, 80,000. Note that all simulated SNPs were in linkage equilibrium and were independent of each other. This situation approximately corresponds to conducting multiple testing after linkage disequilibrium (LD)-based SNP pruning. Uniform distribution of minor allele frequencies was assumed. The disease prevalence was assumed to be 1%. To assess the efficiency with respect to each disease type, we considered two representative disease scenarios, (1) a minimally polygenic disease and (2) a highly polygenic disease, by reference to existing estimates of effect size distributions of disease-associated SNPs for complex diseases [28, 31]. Supplementary Figure 1 shows the distributions for both scenarios. In the minimally polygenic disease that models diseases like rheumatoid arthritis [31], a small proportion of the SNPs (20,000 SNPs) are associated with disease risk and almost all relevant SNPs have large effect sizes. On the other hand, in the highly polygenic disease that models diseases like schizophrenia [28], a large proportion of the SNPs (83,000 SNPs) with modest effect sizes (ORs of <1.05) are associated with disease risk. Genotypes were generated with these settings assuming Hardy–Weinberg equilibrium in the population, and then the allelic association test (χ^2 test with 1 degree of freedom applied to the 2×2 table of case–control allele counts) [7] was conducted for each SNP to obtain summary statistics using the "--assoc" option of PLINK. Although conventional GWAS analysis uses linear or logistic regression to deal with covariates, we did not consider effects of covariates and used the χ^2 test. Under these conditions, we generated 3600 independent simulated data sets and applied the multiple testing strategies to examine

Table 1 The numbers of SNPs declared as significant by each procedure for the two data sets under nominal FWER/FDR level α

α	Bonferroni	BH	ST	ODP	Locfdr	WBF		
						$\pi_1 = 0.01$	$\pi_1 = 0.1$	$\pi_1 = 0.2$
(a) Rheumatoid arthritis								
0.01	125	770	770	718	334	1259	3347	4609
0.05	149	1215	1215	1637	918	2272	7873	12,570
0.10	177	2218	2218	2691	1474	3460	12,329	23,156
0.20	193	3576	3576	4859	2650	5141	24,212	47,359
(b) Schizophreni								
0.01	57	759	776	777	738	822	1746	2494
0.05	128	1560	1693	2207	1683	1352	3853	5964
0.10	156	2939	3221	4889	3086	1826	5884	9665
0.20	211	6050	7112	14,289	7409	2754	10,156	17,954

BH Benjamini–Hochberg, *ST* Storey, *ODP* optimal discovery procedure, *WBF* Wakefield's Bayesian framework

the actual numbers of true and false positives. Then, we empirically estimated the statistical power and the FDR by aggregating these results.

Furthermore, we conducted additional simulations to assess the statistical power with more realistic assumptions of LD structure and the distribution of minor allele frequency (see Section F in the Supplementary Notes for details). The same effect size distributions of disease-related SNPs in the simplistic simulation were considered. We generated genotype counts for SNPs from a multinomial distribution assuming the Hardy–Weinberg equilibrium for genotype frequencies in a general population. We assumed that the physical distance between each SNP was 3 kb and that SNPs in a 30-kb region around an SNP (15-kb upstream and 15-kb downstream) were in LD with the SNP, i.e., 10 SNPs were in LD with the SNP. The strength of the LD and the distribution of allele frequencies were determined based on the actual data derived by the 1000 Genomes Project [32]. Effect sizes of LD SNPs were determined using LD coefficients and allele frequencies based on relevant-related studies of Zondervan and Cardon [33] and Ackerman et al. [34]. Summary statistics were obtained by logistic regression using the generated genotype counts.

Results

Comparison of the number of significant results in analyses of large-scale GWAS data sets

We compared the number of SNPs that were declared as significant by each multiple testing strategy under the nominal FWER/FDR levels for the two GWAS data sets of rheumatoid arthritis and schizophrenia (Table 1). Bonferroni correction detected 149 significant SNPs for the rheumatoid arthritis data set and 128 for the schizophrenia data set at FWER = 5% due to the conservative property of the FWER-controlling approach. On the other hand, the FDR-controlling strategies detected >300 significant SNPs for the rheumatoid arthritis data set and >700 for the schizophrenia data set, even under relatively strict levels (FDR = 1%). In fact, since several hundred additional SNPs were declared as significant with the FDR-based approaches, if the two GWAS studies had used these strategies their conclusions would have been different. Also, the number of SNPs detected using the optimal discovery procedure (ODP) was much larger than with other methods at the same FDR, since the ODP provides the optimal significance ranking for maximizing the expected true positives under fixed expected false positives. Further, Wakefield's Bayesian procedure declared the greatest number of SNPs as significant, but the estimated FDRs were quite different among different prior probability settings, and it is unclear based on these results whether the FDR is accurately controlled by this method (see also "Simulation studies for evaluating actual rates of false positives"). Such inconsistency would confuse geneticists and clinicians, and most practitioners would not be able to determine with certainty how a proper prior probability was selected in practical situations.

Statistical power of the multiple testing strategies for the large-scale GWAS data sets

We evaluated the efficiency of the standard FWER-controlling strategy and the FDR-controlling strategies for the largest data sets of rheumatoid arthritis and schizophrenia that extensively aggregated the currently available GWAS resources. To evaluate the efficiency, we estimated the statistical power of these strategies for both data sets. Figure 1 shows the estimated overall and partial power for

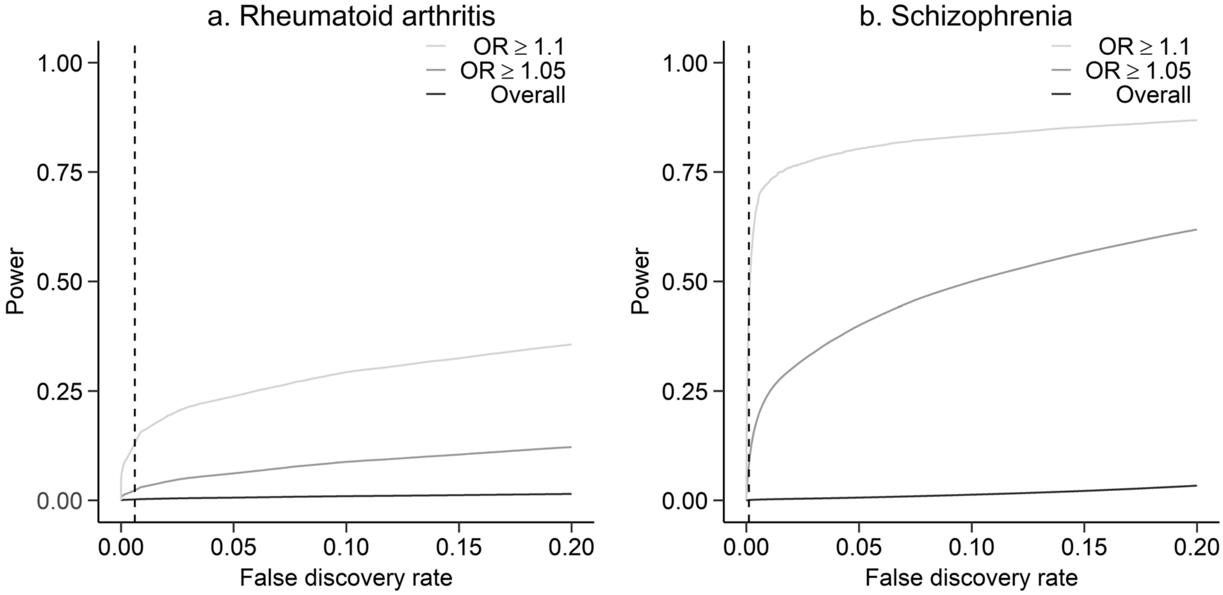

Fig. 1 Plots of estimated overall and partial power (OR ≥ 0.15 and OR ≥ 1.1) for the two data sets. The dashed lines are the estimated FDR levels of the genome-wide significance criterion (0.0061 for rheumatoid arthritis and 0.0011 for schizophrenia). The estimated number of associated SNPs was 238079 (3% of all SNPs) for rheumatoid arthritis and 340221 (27%) for schizophrenia. The estimated numbers of associated SNPs with ORs greater than 1.05 and 1.1 were 21421 (0.34%) and 2290 (0.04%) for rheumatoid arthritis, and 2762 (0.22%) and 390 (0.03%) for schizophrenia, respectively. Note that the number of independently associated SNPs should be much smaller than these estimates since the SNPs are in LD

the two data sets versus estimated FDR levels. These estimated powers can be used to assess the statistical power of the standard FWER-controlling strategy as well as those of the frequentist FDR algorithms [9, 11, 17] in which a P value threshold is used to declare SNPs as significant (see Section D in the Supplementary Notes). The dashed line indicates the estimated FDR level of the genome-wide significance criterion (0.0061 for rheumatoid arthritis and 0.0011 for schizophrenia). At first, the overall power for both data sets was extremely low; the power of the FWER-controlling strategy (using the genome-wide significance threshold [2] of $P < 5 \times 10^{-8}$) was estimated as only 0.09% for the rheumatoid arthritis data set and 0.04% for the schizophrenia data set. These estimates suggest a serious problem, namely that almost none of the disease-related genetic factors would be detected by the conventional strategy even if the largest-scale currently available data sets were used. On the other hand, for the FDR-controlling procedure, the overall power was slightly improved but was still low (less than 1% at FDR = 5%). These results indicate that it will be impossible to detect all disease-related genetic variants if the current GWASs practices are continued, even if significantly more costs and resources are dedicated. The primary reasons for these extremely underpowered results are that most of the disease-related SNPs have modest effects. In Supplementary Figure 2, we present the estimated distributions of effect sizes of disease-related SNPs derived using the empirical Bayes framework (see Section C in the Supplementary Notes), and show that in fact almost all SNPs had modest effect sizes (ORs of <1.05) for these data sets.

On the other hand, the partial power reached relatively high levels, and SNPs with comparably large effect sizes could be successfully detected under reasonable FDR levels; for example, at FDR = 5%, the partial power to detect disease-related SNPs with an OR > 1.1 was estimated to be more than 20% for the rheumatoid arthritis data set and more than 75% for the schizophrenia data set. Such high levels of the partial power compared to the extremely low levels of the overall power might be satisfying for many practitioners, and almost all genetic factors with the largest effects would be detectable under the practically feasible scales of GWASs. Further, these factors might be more clinically and biologically meaningful than others with smaller effects, and most researchers might be implicitly focusing on detecting these factors in current studies.

Simulation studies for evaluating actual rates of false positives

We assessed the actual false positives rates of each multiple testing procedure via simulations. Figure 2 shows the actual FWER/FDR levels estimated from the simulation results versus desired levels. Bonferroni correction, the Benjamini–Hochberg procedure, Storey's procedure, and the ODP properly controlled false positives to nominal

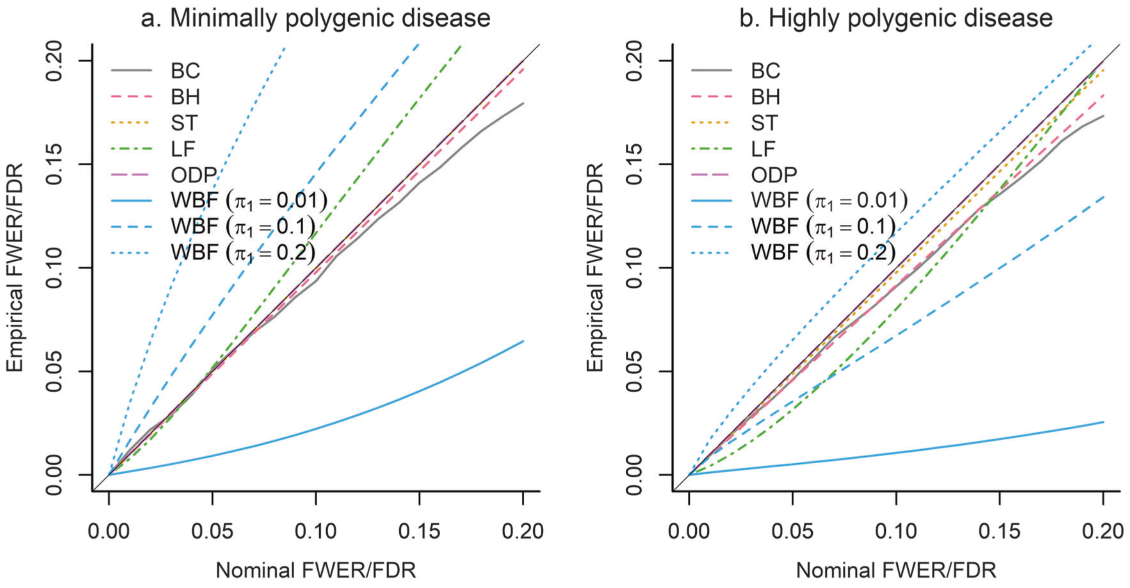

Fig. 2 Actual FWER/FDR levels versus desired levels for both disease scenarios. Empirically estimated FWER are plotted for Bonferroni correction, and estimated FDR are plotted for other procedures. Sample size $N = 40{,}000$ was assumed. The black solid line corresponds to exactly controlled error rates; conservative results are below this line, and liberal results are above this line. BC Bonferroni correction, BH Benjamini–Hochberg, ST Storey, LF Locfdr, ODP optimal discovery procedure, WBF Wakefield's Bayesian framework

FWER/FDR levels. The validity of these frequentist methods for GWASs have been examined previously [13, 14]. However, Wakefield's procedure produced inflated false positives under improper prior probability settings ($\pi_1 = 0.1$ and 0.2 for the minimally polygenic disease, and $\pi_1 = 0.2$ for the highly polygenic disease). This possibly suggests that the Bayesian estimates of the FDR are very sensitive to the prior probability and therefore require careful specification in practical use. Also, the Locfdr procedure failed to properly control false positives. For these reasons, we checked that the marginal distributions of test statistics for the null/non-null components were possibly not accurately estimated via the nonparametric estimation method implemented in this procedure (data not shown). The Locfdr approach might be unstable when applied to GWASs.

Simulation studies for evaluating statistical power for future GWASs

We also conducted large-scale simulation studies to evaluate the efficiency of the multiple testing strategies discussed above, in particular to predict how large the overall/partial power of each strategy would be if the cumulative sizes of GWAS data sets become much larger in future studies. Note that the simulation results of Wakefield's procedure are not shown because the results fluctuated widely due to assumed priors (see "Simulation studies for evaluating actual rates of false positives").

Figure 3 shows plots of empirically estimated overall and partial power versus sample sizes for the two disease scenarios. The FWER-controlling Bonferroni correction required a huge sample size to achieve high overall power levels. In particular, for the highly polygenic disease it would be impossible to achieve sufficient power with the FWER-controlling strategy for practically feasible sample sizes; the estimated overall power of this strategy was only 13% for this scenario, even if 80,000 subjects were aggregated. As suggested by these results, the extremely low power for current available GWAS data sets (less than 0.1%) can be improved by aggregating many more resources, but the levels would still be too low to detect all disease-related genetic factors. On the other hand, the FDR procedures can achieve improved power; for example, the ODP achieved an overall power of 45% for the highly polygenic disease scenario at FDR = 5% with 80,000 subjects. Although it is difficult to detect all disease-related factors regardless of their effect sizes, such a high power level might be satisfying for many practitioners, and these strategies may be useful alternatives to the standard FWER-controlling strategy. Furthermore, the partial power to detect disease-related SNPs with ORs of >1.05 reached >80% for both scenarios using FDR-controlling strategies. These results would provide a completely different perspective compared to the conventional overall power measures in terms of assessing efficiency and designing future studies. Almost all disease-related factors with large effects can be detected by the FDR-controlling strategies if the cumulative sizes of GWAS data sets become much larger, a realistic objective for future studies. Further, the magnitudes of the estimated partial powers for SNPs with ORs of >1.05

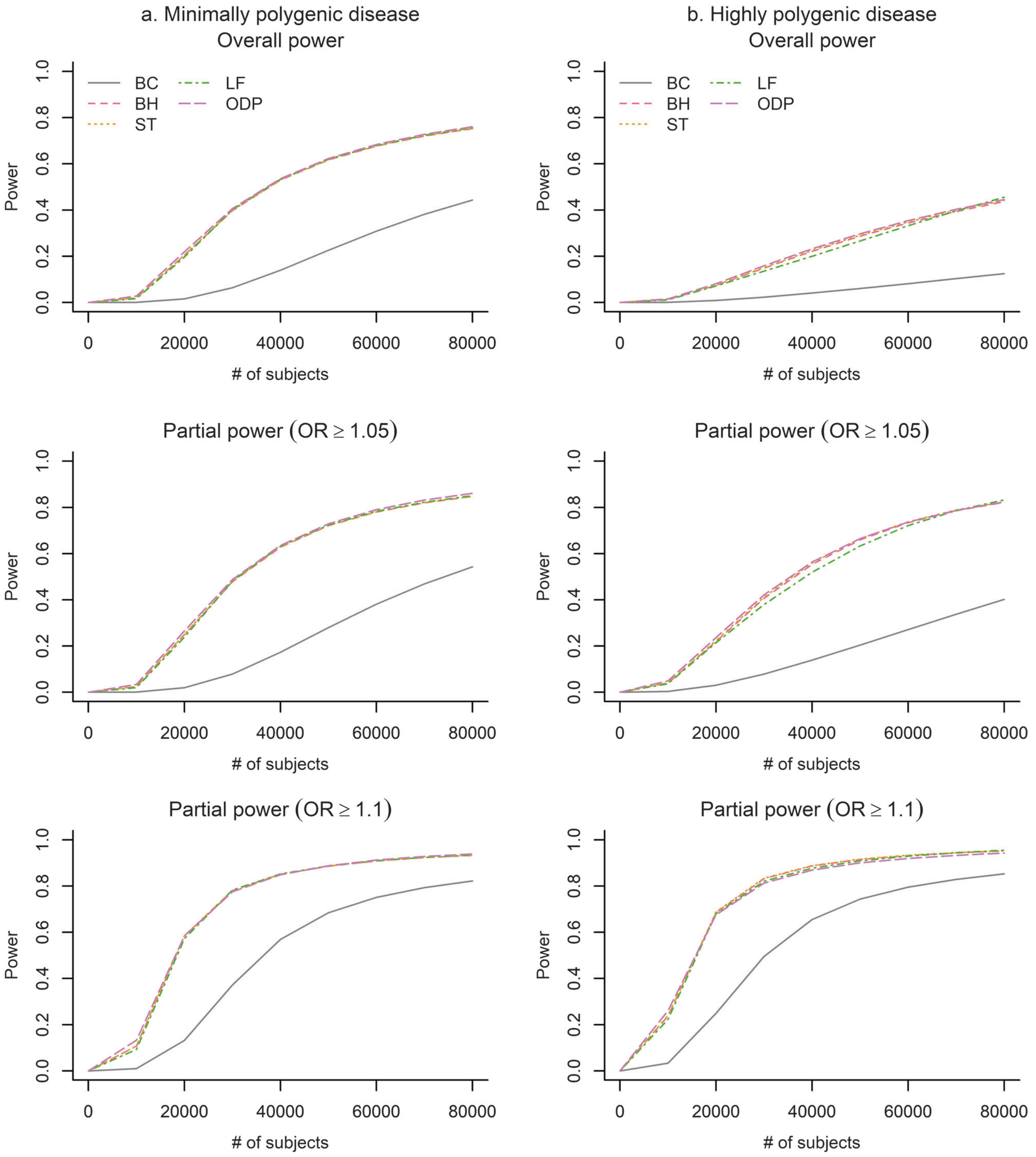

Fig. 3 Empirically estimated overall/partial power versus sample sizes. FWER/FDR level $\alpha = 0.05$ and a million SNPs were assumed. Note that Bonferroni correction is identical to the genome-wide significance criterion under this setting. BC Bonferroni correction, BH Benjamini–Hochberg, ST Storey, LF Locfdr, ODP optimal discovery procedure

using the current GWAS data sets were relatively small (Fig. 1), but these will be sufficiently improved by accumulating samples and almost all clinically and biologically meaningful SNPs will be detected using these strategies.

Table 2 shows a comparison of the overall power for both scenarios under nominal significance levels if 40,000 subjects are assembled. For both scenarios, the overall power of the Bonferroni correction was very low due to its very conservative properties, while the FDR-controlling procedures were more powerful. Also, the overall power of the ODP was consistently much larger at the same FDR than those of other frequentist methods.

Table 2 Overall power for both disease scenarios under nominal FWER/FDR level α and sample size $N = 40{,}000$

α	Bonferroni	BH	ST	ODP	Locfdr
(a) Minimally polygenic disease					
0.01	0.104	0.427	0.428	0.434	0.412
0.05	0.140	0.530	0.531	0.536	0.533
0.10	0.158	0.579	0.580	0.585	0.592
0.20	0.177	0.632	0.634	0.639	0.651
(b) Highly polygenic disease					
0.01	0.033	0.147	0.150	0.155	0.124
0.05	0.041	0.221	0.225	0.232	0.200
0.10	0.045	0.269	0.274	0.281	0.257
0.20	0.050	0.333	0.341	0.348	0.341

BH Benjamini–Hochberg, *ST* Storey, *ODP* optimal discovery procedure

Many practitioners implicitly evaluate the priorities of individual SNPs based on their *P* values, but these results suggest that this is not optimal and that improved power can be achieved by ranking based on the ODP statistics. Table 3 shows a comparison of the partial power for both scenarios under nominal significance levels. High partial power values were achieved compared to the overall power; for example, the Benjamini–Hochberg procedure and Storey's procedure achieved a partial power of >80% at FDR = 1%. However, the ODP showed slightly lower partial power than the other FDR-controlling procedures, perhaps because this procedure maximizes the number of expected true positives regardless of the effect sizes of each relevant SNP, which is not the optimal method for maximizing partial power. For these purposes, alternate methods would be optimal in practice.

Figure 4 shows plots of empirically estimated overall/partial power versus sample sizes for the two disease scenarios under more realistic assumption of LD structure and the distribution of minor allele frequency. Although the estimated statistical powers were lower than the results from the simplistic simulations, the relative ranking of performance of each testing strategy was consistent. The lower power was mainly due to the assumption of the distribution of minor allele frequency; the minor allele frequency of almost all SNPs was assumed to be <0.1 in these simulations instead of the uniform distribution in the simplistic simulations (see Section F in the Supplementary Notes).

Table 3 Partial power (OR ≥ 1.1) for both disease scenarios under nominal FWER/FDR level α and sample size $N = 40{,}000$

α	Bonferroni	BH	ST	ODP	Locfdr
(a) Minimally polygenic disease					
0.01	0.508	0.805	0.805	0.799	0.797
0.05	0.570	0.851	0.851	0.849	0.852
0.10	0.595	0.870	0.871	0.871	0.875
0.20	0.620	0.890	0.891	0.894	0.897
(b) Highly polygenic disease					
0.01	0.607	0.846	0.848	0.828	0.827
0.05	0.655	0.887	0.889	0.869	0.877
0.10	0.675	0.905	0.906	0.888	0.900
0.20	0.694	0.923	0.925	0.909	0.925

BH Benjamini–Hochberg, *ST* Storey, *ODP* optimal discovery procedure

Discussion

In this article, we performed a comprehensive re-assessment of multiple testing strategies to gain insight into overcoming the missing heritability problem from a statistical perspective. In particular, we estimated the statistical power of the current standard FWER-controlling strategies and found that it was extremely low even for the current largest-scale mega-analyses of GWASs. Moreover, it will be difficult to substantially increase power even if many more GWAS resources are aggregated in future studies. Recently published mega-analyses aggregated large-scale resources from a total of >100,000 subjects [4, 6], but the number of relevant genetic variants identified by these studies was much smaller than expected [28, 31]. These results and our power estimates suggest that most disease-related genetic factors will not be identified if the current practices are continued in future studies.

The FDR-based strategies may be promising alternatives, but these have not been used in GWASs, in contrast to their wide use in microarray experiments [16]. One reason is that the proportion of true associations in GWASs is thought to be very small. If this is the case, there is no remarkable difference between the FDR and the FWER because almost all null hypotheses are true (the FDR is exactly equivalent to the FWER if all null hypotheses are true [9]). Another reason is that conventional FDR-controlling procedures cannot appropriately control the actual proportion of false positives within a study when only a few true associations exist [35]. However, as suggested by our efficiency assessments, these strategies might be efficient for GWASs of polygenic traits in which the proportion of true associations is much larger than was expected in past studies. Even though we only considered GWASs with several million SNPs in this study, similar results would be expected for a GWAS with more than 10 million SNPs if the proportion of true associations is as sufficiently large as that in our simulation studies. Each procedure used in this study can be executed within a realistic computation time, even in such large-scale studies. Although one might be concerned about absolute false-positive rates due to the huge number of simultaneous tests in GWASs, these rates are properly

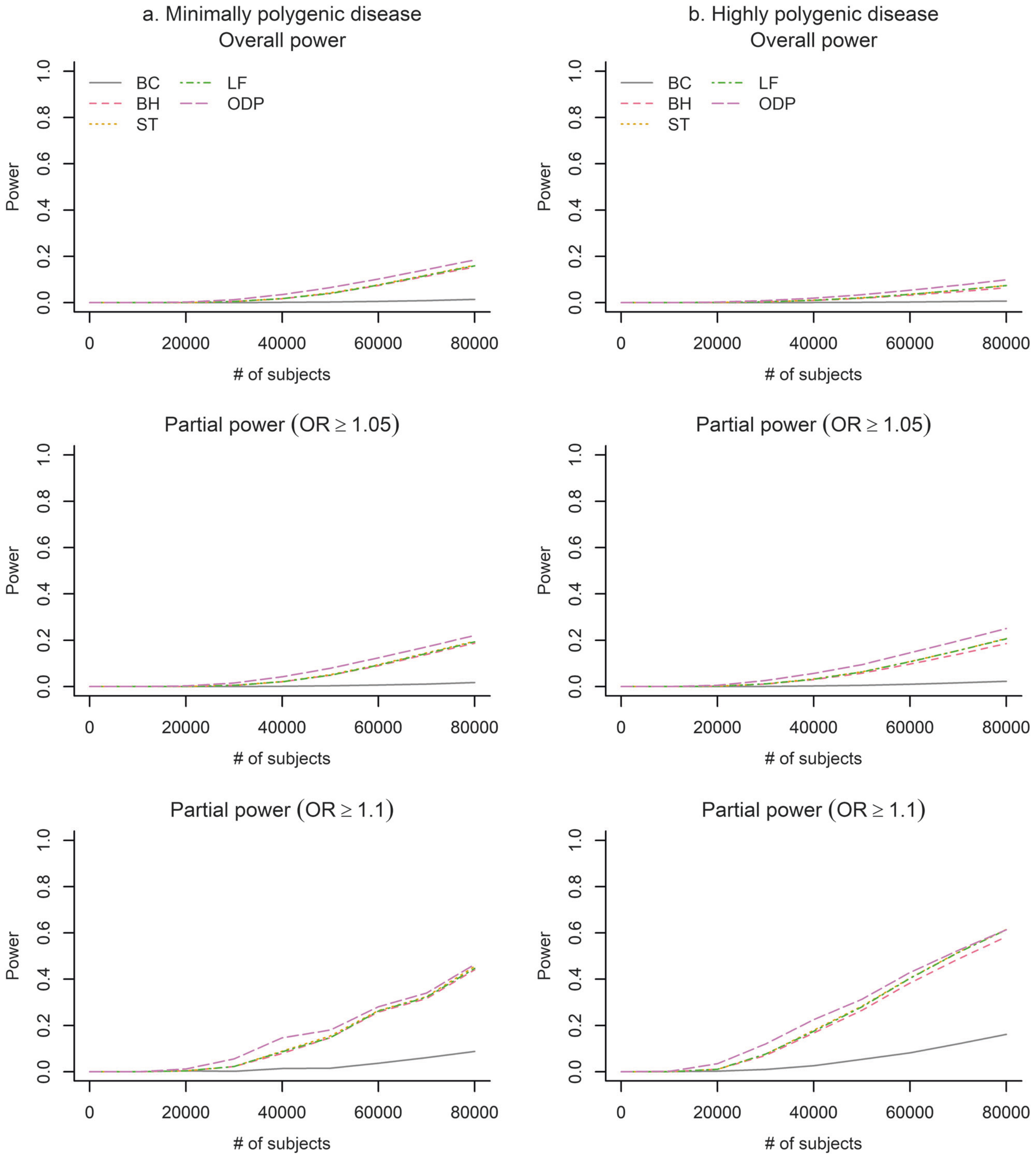

Fig. 4 Overall/partial power versus sample sizes estimated from the simulations with a more realistic genomic property. FWER/FDR level $\alpha = 0.05$ and a million SNPs were assumed. Note that Bonferroni correction is identical to the genome-wide significance criterion under this setting. BC Bonferroni correction, BH Benjamini–Hochberg, ST Storey, LF Locfdr, ODP optimal discovery procedure

controlled by specifying a strict FDR level (for example, FDR < 1%), and the FDR-based strategies would be able to detect a larger number of disease-related variants than the genome-wide significance threshold even under this condition. Also, among the FDR-controlling procedures compared in this study, the ODP showed the best performance in terms of the overall power index, and therefore this procedure should provide the most efficient screening strategies in multi-stage GWASs. In typical multi-stage GWASs, candidate SNPs to be validated in further analysis are usually selected based on a naive *P* value ranking, or on thresholding with a less stringent significance level (for example, $P < 10^{-6}$). These strategies, however, are not optimal in terms of maximizing the number of expected true positives under the fixed number of expected false positives [18, 19]. ODP statistics give the optimal ranking, and should be efficient in the early stage screening of candidate SNPs for further investigations.

The result of the additional simulation study with a more realistic assumption of LD structure and the minor allele frequency distribution showed a loss of statistical power for each strategy, although the FDR-based strategies consistently showed better performance than the FWER-based strategy. The main reason for the loss of power is the assumption of lower minor allele frequency for each SNP in the simulation. Another reason is the assumption of LD structure since all strategies considered in this study work most efficiently if tests are independent, i.e., SNPs are not in LD (see Section B in the Supplementary Notes). On the other hand, the relative ranking of each strategy obtained from the simulation was similar to the ranking from the simplistic simulations, and the ODP consistently showed the best performance. This might be due to the fact that the optimality of the ODP also holds under arbitrary dependence [18] (see Section E in the Supplementary Notes).

However, note that differences of estimated power between each FDR-based strategy were small, although the relative ranking of each strategy was consistent in the simulation studies. These results suggest that the observed relative ranking may not hold for real GWASs. Therefore, in practical applications, validity of assumptions of each procedure on real genomic data should be carefully considered to select the best strategy.

The FDR-based strategies will be able to achieve improved power, and this might be satisfying for practitioners. However, for highly polygenic diseases in which almost all disease-related genetic factors have modest effects, it will be impossible to discover all such factors for practically available sample sizes under a reasonable FDR level. One practical compromise would be to focus on identifying genetic variants with comparably large effects that are clinically meaningful, and the partial power [15] provides a convenient measure for assessing the performance of screening strategies in this context. Our simulation results suggest that the FDR-based strategies can achieve sufficient partial power (>80%) for detecting genetic factors with the largest effects (ORs of >1.05) under practically feasible sample sizes (80,000 subjects), and they may be useful measures for redefining realistic objectives of future GWASs.

In addition, in our simulation studies, we set the effect size distributions by referencing estimates derived from the current large-scale GWASs [28, 31] (Supplementary Figure 1). These studies used the validated estimating method outlined by Stahl et al. [31]. However, there are other plausible biological models, e.g., an L-shaped model, where the frequency of associated SNPs increases as the effect size decreases. Future studies with applicable real data are important to further our understanding. Moreover, the use of partial power is a potentially effective tool for assessing the number of genes that are likely to be clearly identified and have noticeable biological functions. We should further assess these models and tools to determine effective strategies for GWAS and whether choosing to search for all genes, even those with small effects, is worth the significant cost and effort.

As demonstrated in this article, the conventional statistical frameworks for GWASs have limitations, and the missing heritability problem might never be resolved without reconsideration of current practices. Investigators should recognize these facts and possibly change their strategies to overcome this relevant problem.

Acknowledgements This work was supported by Core Research for Evolutional Science and Technology (CREST) from the Japan Science and Technology Agency (Grant no.: JPMJCR1412) and the Practical Research for Innovative Cancer Control from the Japan Agency for Medical Research and Development (Grant no.: 17ck0106266).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42:D1001–D1006.
2. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. Genet Epidemiol. 2008;32:227–34.
3. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.
4. Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014;506:376–81.
5. Ripke S, Sanders AR, Kendler KS, et al. Genome-wide association study identifies five new schizophrenia loci. Nat Genet. 2011;43:969–76.
6. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421–7.
7. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. Basic statistical analysis in genetic case-control studies. Nat Protoc. 2011;6:121–33.
8. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol. 2008;32:381–5.
9. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57:289–300.
10. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29:1165–88.

11. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 2003;100:9440–5.
12. Efron B. Large-scale simultaneous hypothesis testing. J Am Stat Assoc. 2004;99:96–104.
13. Yang Q, Cui J, Chazaro I, Cupples LA, Demissie S. Power and type I error rate of false discovery rate approaches in genome-wide association studies. BMC Genet. 2005;6(Suppl 1):S134.
14. Shi G, Boerwinkle E, Morrison AC, Gu CCC, Chakravarti A, Rao DC. Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS. Genet Epidemiol. 2011;35:111–8.
15. Matsui S, Noma H. Estimating effect sizes of differentially expressed genes for power and sample-size assessments in microarray experiments. Biometrics. 2011;67:1225–35.
16. Crowley J, Hoering A. Handbook of statistics in clinical oncology. 3rd ed. Boca Raton, FL: CRC Press; 2012.
17. Storey JD. A direct approach to false discovery rates. J R Stat Soc Ser B. 2002;64:479–98.
18. Storey JD. The optimal discovery procedure: a new approach to simultaneous significance testing. J R Stat Soc Ser B. 2007;69:347–68.
19. Noma H, Matsui S. The optimal discovery procedure in multiple significance testing: an empirical Bayes approach. Stat Med. 2012;31:165–76.
20. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc. 2001;96:1151–60.
21. Efron B. Microarrays, empirical Bayes and the two-groups model. Stat Sci. 2008;23:1–22.
22. Wakefield J. A Bayesian measure of the probability of false discovery in molecular genetic epidemiology studies. Am J Hum Genet. 2007;81:208–27.
23. Wakefield J. Bayes factors for genome-wide association studies: comparison with P-values. Genet Epidemiol. 2009;33:79–86.
24. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nat Rev Genet. 2009;10:681–90.
25. Jung SH. Sample size for FDR-control in microarray data analysis. Bioinformatics. 2005;21:3097–104.
26. Shao Y, Tseng C-H. Sample size calculation with dependence adjustment for FDR-control in microarray studies. Stat Med. 2007;26:4219–37.
27. Park J-HH, Wacholder S, Gail MH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet. 2010;42:570–5.
28. Ripke S, O'Dushlaine C, Chambert K, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013;45:1150–9.
29. Nishino J, Kochi Y, Shigemizu D, et al. Empirical Bayes estimation of semi-parametric hierarchical mixture models for unbiased characterization of polygenic disease architectures. http://biorxiv.org/lookup/doi/101101/080945 2016.
30. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
31. Stahl EA, Wegmann D, Trynka G, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet. 2012;44:483–9.
32. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.
33. Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. Nat Rev Genet. 2004;5:89–100.
34. Ackerman H, Usen S, Mott R, et al. Haplotypic analysis of the TNF locus by association efficiency and entropy. Genome Biol. 2003;4:R24.
35. Dudbridge F, Gusnanto A, Koeleman BP. Detecting multiple associations in genome-wide studies. Hum Genomics. 2006;2:310–7.