European Journal of Human Genetics (2015) 23, 238–244

www.nature.com/ejhg

ARTICLE

Improving power for robust trans-ethnic meta-analysis of rare and low-frequency variants with a partitioning approach

Sergii Zakharov[1,2], Xu Wang[1], Jianjun Liu[2] and Yik-Ying Teo*[,1,2,3,4,5]

While genome-wide association studies have discovered numerous *bona fide* variants that are associated with common diseases and complex traits; these variants tend to be common in the population and explain only a small proportion of the phenotype variance. The search for the missing heritability has thus switched to rare and low-frequency variants, defined as <5% in the population, but which are expected to have a bigger impact on phenotypic outcomes. The rarer nature of these variants coupled with the curse of testing multiple variants across the genome meant that large sample sizes will still be required despite the assumption of bigger effect sizes. Combining data from multiple studies in a meta-analysis will continue to be the natural approach in boosting sample sizes. However, the population genetics of rare variants suggests that allelic and effect size heterogeneity across populations of different ancestries is likely to pose a greater challenge to trans-ethnic meta-analysis of rare variants than to similar analyses of common variants. Here, we introduce a novel method to perform trans-ethnic meta-analysis of rare and low-frequency variants. The approach is centered on partitioning the studies into distinct clusters using local inference of genomic similarity between population groups, with the aim to minimize both the number of clusters and between-study heterogeneity in each cluster. Through a series of simulations, we show that our approach either performs similarly to or outperforms conventional and recently introduced meta-analysis strategies, particularly in the presence of allelic heterogeneity.
European Journal of Human Genetics (2015) **23,** 238–244; doi:10.1038/ejhg.2014.78; published online 7 May 2014

INTRODUCTION

Meta-analyses of genome-wide association studies (GWAS) have identified hundreds of common genetic variants associated with complex diseases.[1,2] Indeed, meta-analysis of multiple data sets improves the chance of discovering associated variants with moderate or low effect size, which were missed in individual studies due to insufficient power at a stringent genome-wide significance level. Given the development of next-generation sequencing technologies[3] and the fact that for many traits the discovered common genetic variants explain only a small proportion of variability attributable to genetic factors,[4] researchers have recently devoted much attention to rare variants that may hold a clue to the problem of missing heritability. Indeed, evidence that rare variants are associated with complex traits are starting to emerge.[5–7] Given the potential of whole-exome and whole-genome sequencing studies to become as commonplace as GWAS today, it is natural to foresee meta-analysis methods to be applied for the identification of rare variants with moderately larger effect sizes.

Meta-analyses of studies across different populations and ethnic groups have the potential to improve the statistical power to identify rare variants association by increasing the number of samples in the joint analyses.[8] However, when doing trans-ethnic meta-analysis, one faces the problem of effect size heterogeneity, defined as a difference in the effect size across studies, which may stem from: (i) differences in study design (eg, in the definition of the phenotype); (ii) varying impact of genetic variants due to interaction with other variants found at different frequencies across populations; and (iii) different environmental and lifestyle factors. In addition, the analysis of rare variants typically adopts a region-based approach to evaluate the joint genetic burden from multiple variants, given that single-variant statistical methods tend to be underpowered because of stringent multiple testing correction and low allele frequency of individual rare variants. Thus, one also faces the problem of allelic heterogeneity within a region as rare variants are more likely to be population-specific.[9,10] Given these challenges it is important to develop powerful methodologies for trans-ethnic rare-variant meta-analysis that will perform well in the presence of both effect-size heterogeneity and allelic heterogeneity.

In this paper, we introduce a method for performing a region-based trans-ethnic meta-analysis of rare variants that is centered on identifying appropriate partitions of the input data. The method aims to cluster the input studies based on a population genetics argument, before proceeding to measure the association evidence within each cluster. The within-cluster evidence is subsequently combined across clusters to yield a single measurement of statistical evidence of phenotype association for each genomic region. Our method is compared against conventional meta-analysis methodologies, such as those by Fisher and Stouffer, and two of the recently proposed rare-variant meta-analysis methods (MetaSKAT[11] and MV SKAT[12]) with a series of simulations that assumed different extent of

[1]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore; [2]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore; [3]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore; [4]NUS Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore, Singapore; [5]Life Sciences Institute, National University of Singapore, Singapore, Singapore
*Correspondence: Dr Y-Y Teo, Statistics and Applied Probability, National University of Singapore, 6 Science Drive 2 MD3 16 Medical Drive, Singapore 117597, Singapore. Tel: +65 6516 2760; Fax: +65 6872 3919; E-mail: statyy@nus.edu.sg
Received 5 December 2013; revised 20 February 2014; accepted 4 April 2014; published online 7 May 2014

heterogeneity between ancestry groups. The results from our simulations indicated that our method was either comparable to or outperformed existing methodologies for performing trans-ethnic rare-variant meta-analyses.

MATERIALS AND METHODS

Apcluster meta-analysis

Let us assume we have n study groups in a meta-analysis with group-level P-values $p_1,...,p_n$ for a genomic region of interest (P-values are obtained using some region-based rare variants association test). Let us define the study group similarity matrix S as follows:

$$S = \{s_{ij}\}_{i,j=1}^{n} = \left\{\frac{\#(S_i \cap S_j)}{\#(S_i \cup S_j)}\right\}_{i,j=1}^{n},$$

where S_i is the set of all variants (both common and rare) observed in the ith study group, and the operator # maps a set to a number of elements in this set. The numerator thus measures the number of variants that are jointly present in both study i and study j, while the denominator measures the number of variants that at least in one of study i or study j. The similarity measure thus varies between 0 and 1. The intuition behind this similarity measure is based on the population genetics assumption that the shorter the time to the most recent common ancestor (TMRCA) between two populations, the greater the proportion of variants they will share, compared with those populations with a longer TMRCA.

Next, the study groups are partitioned with the use of an affinity propagation clustering algorithm.[13] Let s_{ij} denote the similarity measure between the ith and jth study groups, subsequently known as the ith and jth nodes. The affinity propagation algorithm considers s_{ij} as a measure of how well the jth node is suited to be an exemplar for the ith node (the exemplar is defined as the center of a cluster). *A priori*, all nodes are equally likely to be exemplars. Two kinds of messages are passed between the nodes, namely:

1. The responsibility r_{ij} is sent from the ith node to a candidate jth node, reflecting the evidence for how well the jth node is as an exemplar for the ith node over all other exemplars.
2. The availability a_{ij} is sent from the jth node to the ith node, measuring how appropriately the jth node is chosen by the ith node as an exemplar, relative to the support from other nodes that the jth node is already an exemplar for.

The initial values of all a_{ij} are set to 0, and the algorithm proceeds recursively between the following two updates:

1. Update $r_{ij} = s_{ij} - \max_{j' \neq j}\{a_{ij'} + s_{ij'}\}$;
2. Update $a_{ij} = \min\{0,\ r_{jj} + \Sigma_{i' \notin \{i,j\}}\max\{0,\ r_{i'j}\}$ when $i \neq j$, and $a_{jj} = \Sigma_{j' \neq j}\max\{0,\ r_{j'j}\}$.

At each iteration, the algorithm identifies the jth node as an exemplar for the ith node if $aij + rij$ is maximized, except when the maximum is attained when $j = i$, in which case the ith node itself becomes the exemplar. We utilized the algorithm implemented in the R package 'Apcluster' (http://cran.r-project.org/web/packages/apcluster/index.html), where by default the algorithm terminates if the exemplars have not changed after 100 consecutive iterations, or when the exemplars have not converged after 1000 iterations. When the latter occurs in our implementation, we assume a single cluster containing all the nodes, equivalent to applying the Fisher method to all the studies without partitioning.

Let us denote the obtained clusters of study groups $C_1,...,C_K$. The test statistic is obtained with the following:

(1) within each cluster C_k, for $k \in \{1, 2, \ldots, K\}$, calculate the Fisher test statistic as:

$$\sum_{i \in C_k}(-2\log(p_i))$$

where the corresponding P-value $\hat{p}_k$ can be calculated from a χ^2 distribution at $2|C_k|$ degrees of freedom.

(2) Combine the P-values $\hat{p}_k$ across all K clusters using another round of Fisher meta-analysis:

$$T = \sum_{k=1}^{K} -2\log(\hat{p}_k),$$

which under the null hypothesis is distributed as a χ^2 random variable with $2K$ degrees of freedom. This step assumes that the constituent studies within each cluster are homogeneous and thus the evidence is integrated within a single framework.

Other methodologies for performing rare-variant meta-analyses

We compared our rare-variant meta-analysis setup (Apcluster) with some of the existing approaches: (i) two P-value-based approaches were considered, namely the conventional Fisher combination and the Stouffer inverse standard normal transform method.[14] These two approaches relied on the statistical evidence from two underlying group-level tests: SKAT[15] with the linear kernel and the default beta weights in the R package 'SKAT', and a burden test, which is a likelihood ratio test of a regression coefficient for a collapsed score defined as a number of rare minor alleles within a region of interest born by an individual; (ii) Hom-Meta-SKAT and Het-Meta-SKAT by Lee *et al*[11] for meta-analyzing output from SKAT across studies from homogeneous and heterogeneous ancestry groups respectively, where the R implementation of these methods is available in a package 'MetaSKAT' and the specific function 'MetaSKAT_wZ' was used because we assumed the availability of individual-level data; (iii) the MV SKAT method recently introduced by Hu *et al*,[12] where we only considered the MV method but not the other two approaches (SV-I and SV-E), as these latter methods use additional reference data that would potentially make the comparison of the methods unfair. The authors of MV SKAT also indicated that the MV method performs the best and the performance of the two other approaches (SV-I, SV-E) closely resembled that of MV.[12] We implemented MV SKAT in R because there was no readily available source code by the authors of MV SKAT.

Population genetics simulations for calculating power and false-positive rates

To estimate the type 1 error and power associated with the different methods for performing rare-variant meta-analyses, we performed a series of population genetics simulations with the coalescent simulator *cosi* assuming the best-fit population history model.[16] Specifically, we generated 12 haplotype pools, each containing 10 000 chromosomes across a 1-Mb region, where four of the haplotype pools were generated by random sampling from a background assumed to be of European ancestry, four assumed to be of East Asian ancestry and four of African ancestry. We considered two classes of scenarios, corresponding to the 'non-admixed' and 'admixed' classes (Table 1). In the scenarios under the 'non-admixed' class, the 12 studies in the meta-analysis corresponded to data generated from the 12 haplotype pools (four African studies Af1, Af2, Af3 and Af4; four East Asian studies denoted EA1, EA2, EA3 and EA4; and four European studies denoted E1, E2, E3 and E4). In the scenarios under the 'admixed' class, we generated data for seven studies with different degree of admixture, with three single-ancestry studies conducted assuming Africa (Af), East Asian (EA) and European (E) ancestry respectively, and four studies in admixed populations across the four possible ways of admixture: African and East Asian (Af–EA), African and European (Af–E), East Asian and European (EA–E), and across all three ancestry groups (Af–EA–E). Haplotype pools for the two-ancestry admixed populations were generated by mixing the haplotypes from the respective populations with an admixture proportion that is drawn independently for each data replicate from a Uniform (0.2, 0.8) distribution. To generate the admixture proportions for Af–EA–E, we sampled a random vector $v = (v_1, v_2, v_3)$ uniformly from the simplex $\{(x_1, x_2, x_3)$: $x_1 + x_2 + x_3 = 1,\ x_i \geq 0.2,\ i = 1,2,3\}$ according to the algorithm presented by Onn and Weissman.[17]

Following Wu *et al*,[15] we simulated 1000 data replicates of 30-kb region each under the assumption of (i) the null hypothesis of no association with a

quantitative phenotype to calculate the false positive rate and (ii) the alternative hypothesis where the region carried variants that are functionally associated with the phenotype to calculate power.

To calculate power, we randomly chose 5% of the L rare variants (each with observed minor allele frequency $\leq$ 1%) that existed within the corresponding haplotype pool in each study to be causal. Note that this meant different rare variants may be selected to be causal in the different studies; thus, our simulation setup naturally models the situation of allelic heterogeneity. For the purpose of presentation, let us enumerate the causal variants for a study population by the index range $1 \leq l \leq L'$ (with $L' = \text{round}[0.05L]$) and other rare non-causal variants by $L'+1 \leq l \leq L$. We randomly sampled 2000 chromosomes from the respective haplotype pool to form the genotype data for 1000 individuals, and generated a quantitative trait y_i for the ith individual as $y_i = \sum_{l=1}^{L'} b_l g_{il} + e_i$ with $i = 1,...,1000$, b_l and g_{il} are the effect size and minor allele count at the genotype for the ith individual at the lth causal variant (for $l = 1,...,L'$), respectively, and e_i follows a standard normal distribution. Following Wu *et al*,[15] we assigned $b_l = 0.4|\log_{10} MAF_l|$, where MAF_l is a minor allele frequency of the lth causal variant calculated from a respective haplotype pool. To simulate the quantitative trait under the null model of no association, y_i is simply obtained by sampling from a standard normal distribution.

Across the two classes that assumed no admixture and admixed populations respectively, we considered different scenarios of effect size heterogeneity (Table 1). Under the class of 'non-admixed' populations, we considered five scenarios where the simulated genomic region is associated: (1) in all twelve studies (N12); (2) only in the non-African studies (N8); (3) only in the European studies (N4); (4) in two African, three East Asian and three European studies (compound heterogeneity, abbreviated N8C); and (5) in one African, one East Asian and two European studies (N4C). Under the class of 'admixed' populations, we considered three scenarios where the simulated region is associated: (1) in all the seven studies (A7); (2) only in the studies with at least some European ancestry (A4); and (3) only in European and East-Asian studies (A2).

The power is calculated from the 1000 data replicates in each scenario as the proportion of data replicates with meta-analysis P-value $< 10^{-6}$. To model the false-positive rate when the significance threshold was kept at 10^{-6}, we adopted the procedure described by Lee *et al*[11] to avoid the computationally intensive manner of performing at least 10^7 iterations: we simulated 2000 genomic regions under each scenario and 10 000 phenotype replicates under the null model for each genomic region. This approach provided 2×10^7 P-values to estimate the empirical Type 1 error associated with a significant threshold of 10^{-6}.

For P-value-based methods the two group-level tests (SKAT and burden test) were applied to variants with MAF $\leq$ 1% in a data replicate. For Hom-Meta-SKAT and Het-Meta-SKAT, if a variant was rare in one population but not rare (MAF $>$1%) in another population group, we treated this variant as missing for all the study groups for which it was found to be not rare. This procedure ensured a fair comparison between the P-value-based meta-analysis methods and those developed by Lee *et al*.[11]

Theoretical power simulations assuming non-central χ^2 distributions

In addition to population genetics simulations, we considered a theoretical model to evaluate the performance of the Apcluster approach for combining P-values against the Fisher approach. We first assumed that, without loss of generality, the test statistic for each study under the null hypothesis was distributed as a χ^2 distribution with 1 degree of freedom, whereas under the alternative hypothesis, the test statistic followed a non-central χ^2 distribution with a non-centrality parameter α. Instead of simulating genetic and phenotypic data, we can simulate the resultant test statistics of the association analyses directly from the distributions under the null and alternative hypotheses according to the eight simulation scenarios that we assumed for the non-admixed and admixed classes (see Table 1). For example, under the N8 scenario, eight test statistics will be drawn from a non-central χ^2 distribution with a non-centrality parameter α while the remaining four test statistics will be drawn from a χ_1^2 distribution. The P-values corresponding to these test statistics are obtained after mapping against the quantiles of a χ_1^2 distribution. The power of the Fisher and Apcluster methods is derived as the proportion of 5000 iterations where the meta-analysis P-value is more significant than 10^{-6}. For the Apcluster partitioning, we combined the P-values by mimicking the study partitionings that were most often observed in our population genetics simulations in the respective scenarios (see Supplementary Table S1)—in the non-admixed case, the African, East Asian and European studies were always clustered by ancestry and thus we partitioned our test statistics accordingly; in the admixed case, we selected the two most common partitionings to consider in our theoretical considerations, denoted as P1 and P2 (Table 2).

Apcluster method website

The R functions for the Apcluster method is packaged together with a sample data set and accompanying codes for performing the association analysis. The simulated data sets for comparing power and the R package can be downloaded from http://www.statgen.nus.edu.sg/~software/apcluster.html.

RESULTS

Type 1 error rates

We evaluated the false-positive rate of our approach (Apcluster) for combining P-values against two classical strategies: (i) the Fisher method and (ii) the Stouffer inverse-normal method, with 20 000 000 simulations under the null hypothesis of no association. At a significance threshold of 10^{-6}, we observed that the empirical false-positive rates of the Apcluster approach when applied to the output of SKAT ranged between 4.5×10^{-7} and 5.5×10^{-7} for the two classes of simulations that assumed 12 populations without admixture and 7 populations with some admixture, respectively (Table 3). These figures were comparable to those obtained by the Fisher method (3.0×10^{-7}, 6.5×10^{-7}) and by the Stouffer method (5.0×10^{-7}, 6.5×10^{-7}). Similar results were observed with the output from the burden analyses, with empirical false-positive rates of 1.9×10^{-6} and

Table 1 Scenarios considered in population genetics simulations

Scenario annotation	*Simulated population heterogeneity*[a]	*Studies in meta-analysis*[b]	*Studies under the alternative hypothesis*
N12	Non-admixed	Af1–Af4, EA1–EA4, E1–E4	All
N8	Non-admixed	Af1–Af4, EA1–EA4, E1–E4	EA1–EA4, E1–E4
N4	Non-admixed	Af1–Af4, EA1–EA4, E1–E4	E1–E4
N8C	Non-admixed	Af1–Af4, EA1–EA4, E1–E4	Af1–Af2, EA1–EA3, E1–E3
N4C	Non-admixed	Af1–Af4, EA1–EA4, E1–E4	Af1, EA1, E1, E2
A7	Admixed	Af, EA, E, Af–EA, Af–E, EA–E, Af–EA–E	All
A4	Admixed	Af, EA, E, Af–EA, Af–E, EA–E, Af–EA–E	E, Af–E, EA–E, Af–EA–E
A2	Admixed	Af, EA, E, Af–EA, Af–E, EA–E, Af–EA–E	EA, E

[a]Under the non-admixed case, all studies were simulated uniquely from one ancestry group without admixture, whereas under the admixed case, three studies were simulated uniquely from one ancestry group without admixture, and the remaining four studies were simulated considering all possible ways of admixture between three ancestry groups.
[b]Indicative of the ancestry group in which the study was simulated in, consisting of African (Af), East Asian (EA) and European (E). Study symbols with hyphens indicate the ancestries contributing to the admixture, where Af–EA indicates a study simulated with admixed samples of African–East Asian ancestry.

1.2×10^{-6} by Apcluster for the two corresponding classes, compared with those by the Fisher method (1.9×10^{-6}, 1.3×10^{-6}) and the Stouffer method (1.6×10^{-6}, 1.2×10^{-6}). There was no significant elevation of false positives by these three approaches. We did not evaluate the empirical Type 1 error rates of Hom-Meta-SKAT, Het-Meta-SKAT and MV SKAT as the computation burden of these methods were extremely high for the number of simulations we performed, but evidence from the original publications of these methods indicated that these approaches are well calibrated to control Type 1 error.[11,12]

Table 2 Direct simulation from distributions of test statistics

Scenario annotation	*Number of studies under H0*	*Number of studies under H1*	*Apcluster partitioning*[a]
N12	0	12	(4A), (4A), (4A)
N8	4	8	(4N), (4A), (4A)
N4	8	4	(4N), (4N), (4A)
N8C	4	8	(2A, 2N), (3A, 1N), (3A, 1N)
N4C	8	4	(1A, 3N), (1A, 3N), (2A, 2N)
A7	0	7	P1: (3A), (4A) P2: (A), (2A), (4A)
A4	3	4	P1: (2A, 1N), (2A, 2N) P2: (1N), (2A), (2A, 2N)
A2	5	2	P1: (2A, 1N), (4N) P2: (1A), (1A, 1N), (4N)

[a]Indicates the number and constituents of partitions observed from the population genetics simulations for the respective scenario, where the constituents refer to whether the study was under the null hypothesis (N) or under the alternative hypothesis (A). An example partition of {(2A, 2N), (3A, 1N), (3A, 1N)} indicates that three clusters were observed, where the first cluster consisted of two studies under the alternative, and two studies under the null; the second and third clusters each consisted of three studies under the alternative and one under the null.

Table 3 Empirical Type 1 error rates of *P*-value-based meta-analysis methods at a significance threshold of *P*-value $< 10^{-6}$

Number of studies	*Fisher SKAT*	*Stouffer SKAT*	*Fisher burden*	*Stouffer burden*	*Apcluster SKAT*	*Apcluster burden*
12	3.0×10^{-7}	5.0×10^{-7}	1.9×10^{-6}	1.6×10^{-6}	4.5×10^{-7}	1.9×10^{-6}
7	6.5×10^{-7}	6.5×10^{-7}	1.3×10^{-6}	1.2×10^{-6}	5.5×10^{-7}	1.2×10^{-6}

Power comparisons

In evaluating the power of the different meta-analytic procedures, we chose 5% of the rare variants in each region to be functional and assigned the values of a quantitative trait as a function of the number of copies of the causal alleles carried. Power was then estimated as the proportion of the 1000 simulations where the specific meta-analysis approach yielded statistical evidence of *P*-value $< 10^{-6}$. We considered two classes of scenarios in the simulations involving populations from three ancestry groups mimicking Africans, East Asians and Europeans. The first class assumes scenarios where studies originated from populations in these three ancestry groups without any admixture, whereas the second class assumes scenarios where studies originated from populations that are admixed between these three ancestry groups with different degrees of admixture.

We observed that the burden tests yielded lower power in all the scenarios in our simulations when meta-analyzed with *P*-value-based methods (Figure 1). This was unsurprising as the burden test was expected to perform well when a high proportion of rare variants were associated with the outcome, whereas our simulations only assumed 5% of the rare variants to be causal. Our simulations assumed the presence of allelic heterogeneity across the different ancestry groups, which probably explained why MV SKAT and Hom-Meta-SKAT yielded consistently poorer performance than Het-Meta-SKAT. Intriguingly, we observed the Fisher method in combining the output from SKAT consistently outperformed

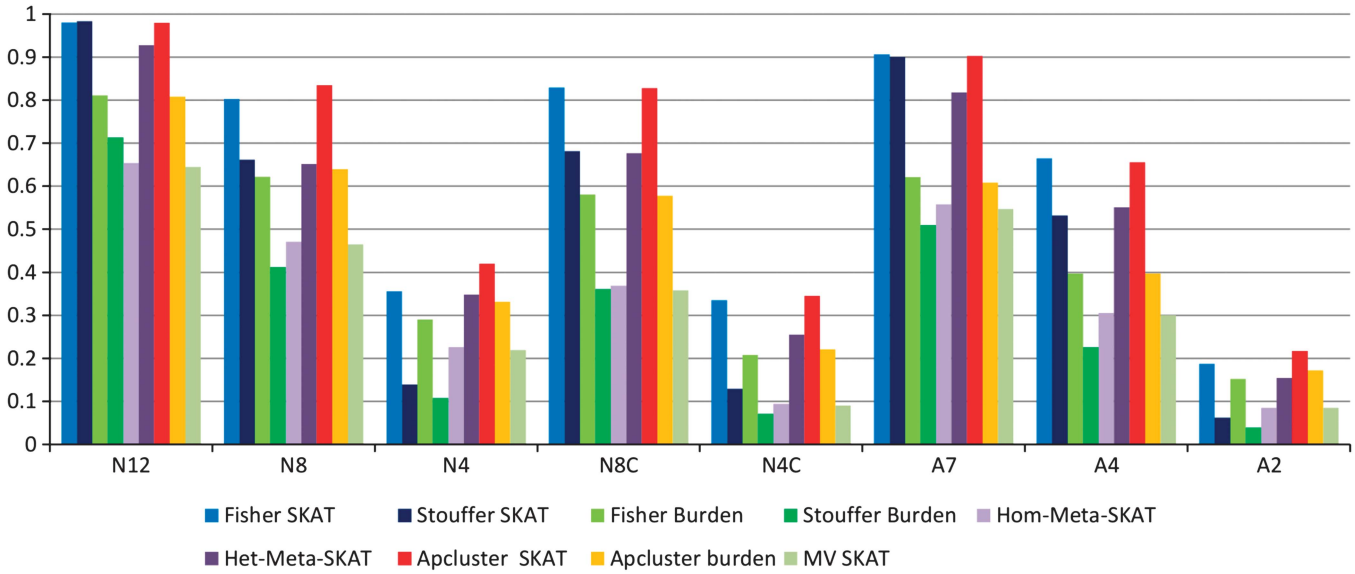

Figure 1 Empirical power of different methods to perform meta-analysis. Empirical power to identify genomic regions simulated with 5% of the rare variants to be associated with a quantitative phenotype, as calculated from 1000 iterations. We consider eight different scenarios corresponding to different degrees of allelic heterogeneity and admixture between populations from three ancestry groups between 12 studies (in scenarios N12, N8, N4, N8C and N4C) and 7 studies (in A7, A4 and A2). Two approaches (burden test, SKAT) were used to perform the association analysis within each study prior to meta-analyzing with Apcluster, the Fisher method or Stouffer inverse-normal method. We additionally included three recently introduced methods: MV SKAT, Hom-Meta-SKAT and Het-Meta-SKAT.

Het-Meta-SKAT, a finding that contradicted the results of Lee *et al.*[11] This remained so even when we attempted to reproduce their findings by limiting the simulations to only three study groups under the alternative hypothesis for 3-kb genomic regions with rare variants defined as those with MAF $\leq$ 3% (Supplementary Table S2, Supplementary Figures S1 and S2), although we did replicate their finding that Hom-Meta-SKAT was superior to Fisher SKAT when we assumed allelic homogeneity between the populations.

Our proposed approach with the output from SKAT (Apcluster SKAT) yielded either comparable or higher power to Fisher SKAT, where the largest gain in power were achieved in the N8, N4 and A2 scenarios with 3.2%, 6.4% and 3% increase in power, respectively, compared with the next best method respectively. These three scenarios corresponded to the situations where there were greater degrees of heterogeneity between the 12 non-admixed and 7 admixed studies. To explore the origins of the power gain, we investigated the partitioning of the studies as derived by the affinity propagation clustering algorithm. We observed that in the non-admixed class, all 12 studies were always clustered into three partitions each containing four studies from an ancestry group. In the N8 scenario, this partitioning separated the four African studies under the null hypothesis from the other eight non-African studies under the alternative hypothesis, and presented a meta-analysis framework with only six degrees of freedom compared with the traditional 24 degrees of freedom.

In addition, we have performed simulations for the non-admixed scenarios where we assumed allelic homogeneity for studies from the same ancestry, but allelic heterogeneity for studies from different ancestries. Specifically, the same causal variants were assumed for studies originating from populations in the same ancestry. Our results indicated that while the methods that explicitly relied on allelic homogeneity (Hom-Meta-SKAT, MV SKAT) produced higher power in the N12, N8 and N4 scenarios than what was observed in the previous set of simulations, Apcluster always yielded performance comparable to these methods (Supplementary Figure S3). However, in scenarios that assumed heterogeneity in the effect sizes (N8C and N4C), the former two methods yielded significantly lower power than Apcluster.

Theoretical power simulations

We simulated the test statistics directly from corresponding probability distributions under the null and alternative hypotheses according to the eight scenarios we have assumed in our population genetics simulations, and combined the *P*-values obtained from mapping the test statistics against a χ_1^2 distribution with the Fisher method and Apcluster. Overall, the trend in the power difference between the two approaches was independent of the non-centrality parameter α, although the actual magnitude of the power difference depended on the size of α. We observed that the Fisher method yielded marginally higher power for the N12, A7 and A4 scenarios, where the power difference was never more than 5% even at the most penalizing α (Figure 2a, d and e). For scenarios N12 and A7, all of the test statistics were simulated under the alternative hypothesis, and thus the partitioning by Apcluster did not contribute any useful information to minimize heterogeneity in terms of association evidence. However, Apcluster yielded substantially higher power in scenarios N8, N4 and A2 (Figure 2b, c and f) with a potential power gain of 4.3, 16.7% and 13% in the N8, N4 and A2 scenarios, respectively, because the partitioning by Apcluster separated the studies simulated under the null and alternative hypotheses. The powers of Apcluster and Fisher method were around the same when the partitioning was less informative, such as when each partition contained studies simulated under both the null and alternative hypotheses in the compound heterogeneity scenarios (N8C, N4C). The two most frequently observed partitionings of the seven studies (P1, P2) in the A2 model both yielded higher power than the Fisher method, because four out of the five studies under the null hypothesis were clustered together in both P1 and P2, thus significantly reducing the degrees of freedom for the test statistic under the null hypothesis.

DISCUSSION

We have introduced a simple procedure for meta-analyzing genetic association studies of rare and low-frequency variants. Our approach partitions the studies into distinct clusters according to the extent of similarity surrounding the locality of each genomic region, before combining the evidence from standard burden-based or region-based analysis of rare variants with the Fisher method. We show from our simulations that this additional step of partitioning increases the power to identify regions that contain variants that are genuinely associated with a phenotype, without elevating the false discovery rate. In addition, our method is robust to the presence of allelic heterogeneity across studies, especially when the pattern of allelic heterogeneity correlates with background genomic heterogeneity.

The intuition behind our method is straightforward: the meta-analysis of studies under the null hypothesis yields a measure of statistical evidence that is no different from a single study under the null hypothesis, but the conventional meta-analysis wastes valuable degrees of freedom in accounting for the multiple studies, whereas the meta-analysis of studies under the alternative hypothesis yields substantially stronger evidence against the null hypothesis, and clustering only such studies for a joint analysis avoids attenuating the evidence that can be brought about when studies under the null are included. In designing our method to cluster the studies, we have leveraged on the principle that rare and low-frequency functional variants have the tendency to segregate according to ancestry, and this appears to be a reasonable assumption in light of recent reports from whole-genome sequencing at the population level.[18,19] This is very much akin to the approach utilized in MANTRA that groups populations by F_{ST}.[20]

Our Apcluster approach provides a framework to perform the meta-analysis, although it relies on association evidence produced by methods such as SKAT or the burden test. The approach can naturally be applied to new methods for analyzing associations. The current setup assumes that study-level information such as the list of variants present within each study is available, in order to derive the extent of sharing between studies for every genomic locus to identify the partitions. In this manner, the partitionings obtained may differ according to the region under consideration and in principle provide a more informative manner to represent the localized genetic diversity between the population groups. Sharing study-level information is straightforward because no individual-level information is exchanged. However, this is not strictly necessary and prior information on the degree of relatedness between study groups can be incorporated to cluster the studies.

We emphasize that our method provides an additional scheme to discover phenotype associations, especially in light of the stringent criterion in defining genome-wide significance where including a few studies in a naive meta-analysis can attenuate the overall signal below the threshold of significance. Findings that emerge from this scheme need to be subject to the same scrutiny and need further validation just as regions that emerge from the conventional Fisher method or other more sophisticated approaches. The partitioning of the studies

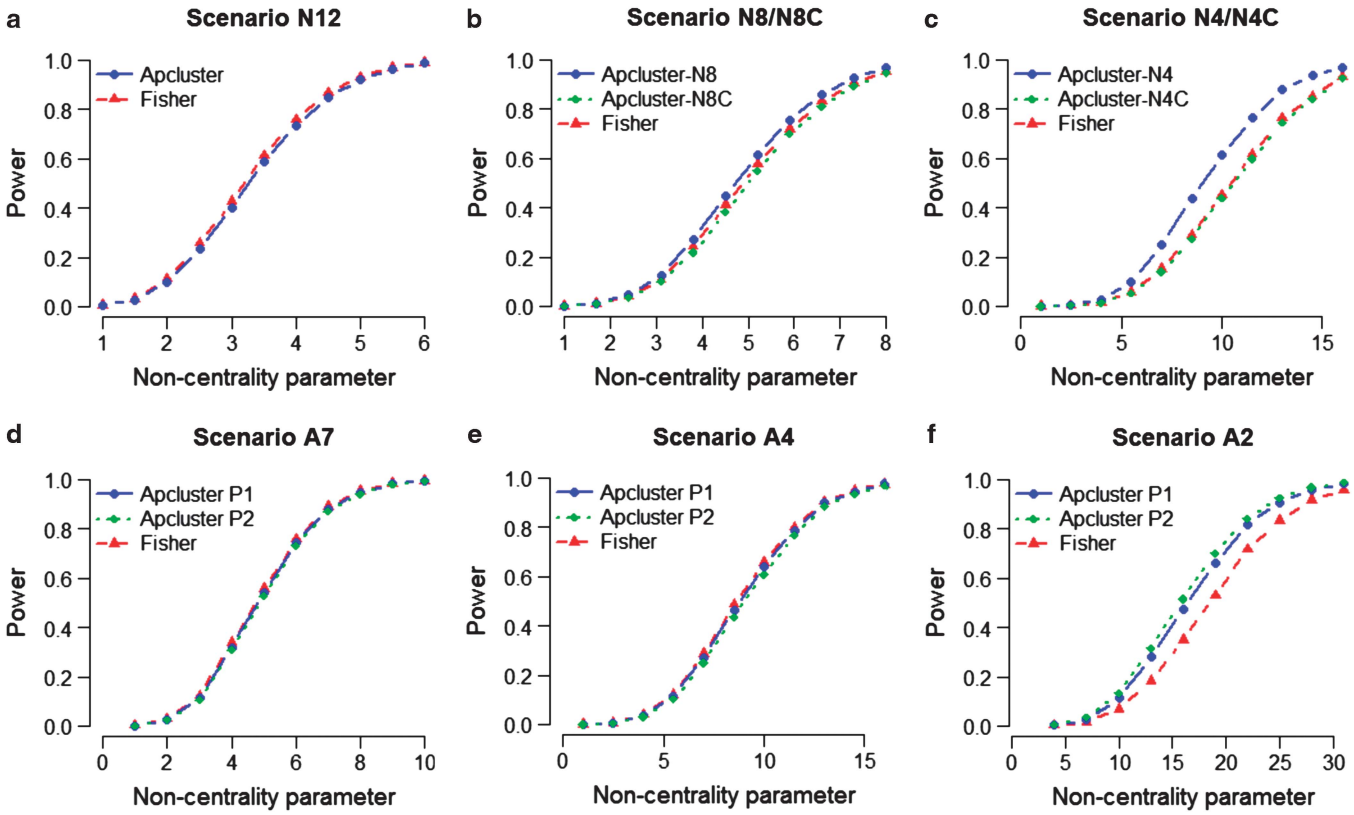

Figure 2 Power simulations from theoretical distributions. Under the alternative hypothesis, the test statistics of the association analysis follow a non-central χ^2 distribution with a specific non-centrality parameter, while the *P*-value is calculated from a central ${\chi_1}^2$ distribution. By sampling directly from the distributions under the null and alternative hypotheses, the power of Apcluster and the Fisher method for combining *P*-values can be evaluated across different non-centrality parameters (horizontal axes) for the eight scenarios that corresponded to different degrees of allelic heterogeneity and admixture. Scenarios for panels **a**–**f** are similar to those assumed in our population genetics simulations. Scenarios N8 and N8C are jointly represented in the same panel, as with scenarios N4 and N4C in a separate panel, since each of the two pairs corresponded to the same number of studies simulated under the null and alternative hypotheses respectively. For the admixed populations, the two most frequently observed partitionings (corresponding to P1 and P2) of the populations in the population genetics analyses were considered in the power simulations directly from the theoretical distributions. The range of non-centrality parameter for each scenario is chosen such that the power spectrum spans between 0 and 100%.

may guide the selection of the populations in which the replication experiment can be performed , as there may be natural population or ancestry clades that are carrying the association signals. At the very least, it can help to put in perspective the failure to reproduce a finding, if validation was carried out in populations that coincided with those studies in a cluster that yielded evidence in favor of the null hypothesis.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

SZ is a recipient of the Singapore International Graduate Award and acknowledges support of the Agency for Science, Technology and Research (A*STAR) Singapore. XW and Y-YT acknowledge support from the Saw Swee Hock School of Public Health from the National University of Singapore and the National Research Foundation Singapore (NRF-RF-2010-05).

1 Teslovich TM, Musunuru K, Smith AV *et al*: Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010; **466**: 707–713.
2 Zeggini E, Scott LJ, Saxena R *et al*: Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008; **40**: 638–645.
3 Shendure J, Ji H: Next-generation DNA sequencing. *Nat Biotechnol* 2008; **26**: 1135–1145.
4 Cirulli ET, Goldstein DB: Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010; **11**: 415–425.
5 Green EK, Grozeva D, Sims R *et al*: DISC1 exon 11 rare variants found more commonly in schizoaffective spectrum cases than controls. *Am J Med Genet B Neuropsychiatr Genet* 2011; **156**: 490–492.
6 Ramagopalan SV, Dyment DA, Cader MZ *et al*: Rare variants in the CYP27B1 gene are associated with multiple sclerosis. *Ann Neurol* 2011; **70**: 881–886.
7 Xie P, Kranzler HR, Krauthammer M *et al*: Rare nonsynonymous variants in alpha-4 nicotinic acetylcholine receptor gene protect against nicotine dependence. *Biol Psychiatr* 2011; **70**: 528–536.
8 Morris AP: Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* 2011; **35**: 809–822.
9 Nelson MR, Wegmann D, Ehm MG *et al*: An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 2012; **337**: 100–104.
10 Tennessen JA, Bigham AW, O'Connor TD *et al*: Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 2012; **337**: 64–69.
11 Lee S, Teslovich Tanya M, Boehnke M, Lin X: General framework for meta-analysis of rare variants in sequencing association studies. *Am J Hum Genet* 2013; **93**: 42–53.
12 Hu Y-J, Berndt Sonja I, Gustafsson S *et al*: Meta-analysis of gene-level associations for rare variants based on single-variant statistics. *Am J Hum Genet* 2013; **93**: 236–248.
13 Frey BJ, Dueck D: Clustering by passing messages between data points. *Science* 2007; **315**: 972–976.
14 Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RM: Adjustment during army life. *The American Soldier*. Princeton, NJ, USA: Princeton University Press, 1949; Vol 1.
15 Wu Michael C, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011; **89**: 82–93.

16 Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D: Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 2005; **15**: 1576–1583.
17 Onn S, Weissman I: Generating uniform random vectors over a simplex with implications to the volume of a certain polytope and to multivariate extremes. *Ann Oper Res* 2011; **189**: 331–342.
18 Mathieson I, McVean G: Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 2012; **44**: 243–246.
19 Abecasis GR, Auton A, Brooks LD *et al*: An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56–65.
20 Morris AP: Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* 2011; **35**: 809–822.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)