Published in final edited form as:
Genet Epidemiol. 2009 December ; 33(8): 700–709. doi:10.1002/gepi.20422.

Pathway analysis by adaptive combination of P-values

Kai Yu[1,*], **Qizhai Li**[1,2], **Andrew W. Bergen**[3], **Ruth M. Pfeiffer**[1], **Philip S. Rosenberg**[1], **Neil Caporaso**[1], **Peter Kraft**[4], and **Nilanjan Chatterjee**[1]

[1]Division of Cancer Epidemiology and Genetics, NCI, Rockville MD, USA

[2]Academy of Mathematics and Systems Science, CAS, Beijing, China

[3]Molecular Genetics Program, Center for Health Sciences, SRI International, Menlo Park, CA, USA

[4]Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA

Abstract

It is increasingly recognized that pathway analyses—a joint test of association between the outcome and a group of single nucleotide polymorphisms (SNPs) within a biological pathway—could potentially complement single-SNP analysis and provide additional insights for the genetic architecture of complex diseases. Building upon existing P-value combining methods, we propose a class of highly flexible pathway analysis approaches based on an adaptive rank truncated product (ARTP) statistic that can effectively combine evidence of associations over different SNPs and genes within a pathway. The statistical significance of the pathway-level test-statistics is evaluated using a highly efficient permutation algorithm that remains computationally feasible irrespective of the size of the pathway and complexity of the underlying test-statistics for summarizing SNP- and gene-level associations. We demonstrate through simulation studies that a gene-based analysis, that treats the underlying genes, as opposed to the underlying SNPs, as the basic units for hypothesis testing, is a very robust and powerful approach to pathway-based association testing. We also illustrate the advantage of the proposed methods using a study of the association between the nicotinic receptor pathway and cigarette smoking behaviors.

Keywords

Pathway analysis; genetic association study; permutation procedure

Introduction

With the advance in high-throughput genotyping technology, genome-wide association studies (GWAS) have emerged as an efficient approach to identifying common polymorphisms underlying complex diseases or traits [Consortium 2007; Hunter, et al. 2007; Yeager, et al. 2007]. The main goal of GWAS is to identify and replicate single nucleotide polymorphisms (SNPs) that are associated with a complex trait. In other words, the testing unit often is a SNP (called single-SNP analysis). It is increasingly recognized that pathway analysis—a joint test of association between the outcome and a group of SNPs or genes within a defined biological pathway--might complement single-SNP analysis and provide additional insights about the genetic architecture of complex diseases [Wang, et al. 2007]. If, for example, multiple SNPs in a pathway contribute to disease susceptibility, but individually each SNP has a relatively

*To whom reprint requests should be sent at Division of Cancer Epidemiology and Genetics, National Cancer Institute, EPS 8050, MSC 7244, 6120 Executive Blvd., Rockville, MD 20892; Phone: 301-496-4153; Fax:301-402-0081; E-mail: yuka@mail.nih.gov.

small effect, the evidence for an association between an individual SNP and the outcome could be too weak to be detected by the single-SNP analysis. In contrast, pathway analysis offers the opportunity to combine association evidence from multiple genetic variants and thus potentially has a better chance of identifying the association between the pathway and the disease.

One approach to summarizing the association evidence across SNPs within a biological pathway is to combine SNP-level P-values using a P-value combining method, such as the Fisher product method [Fisher 1932], the truncated product method that uses the product of all P-values at less than a pre-selected threshold as the summary statistic [Zaykin, et al. 2002], or the rank truncated product (RTP) method, which uses the product of the *K* most significant P-values as the summary statistic [Dudbridge and Koeleman 2003]. We call this type of pathway analysis the SNP-based pathway analysis. A potential weakness of the SNP-based approach is that it ignores the structure of the underlying candidate genes that define the biologic pathway. Given that different genes can have different sizes, SNP density and linkage disequilibrium (LD) patterns, a SNP-based approach may not be able to properly weigh the association evidence from different genes that are the natural biologic units of association.

In this article, we propose a gene-based pathway analysis approach that combines gene-level association evidence through an adaptive rank truncated product (ARTP) method [Dudbridge and Koeleman 2004; Hoh, et al. 2001]. This gene-based approach consists of the following two steps: 1. obtain the standardized summary for the association evidence between a gene and the outcome, e.g., based on the multiple-testing adjusted P-value associated with the most significant SNP within a gene; 2. use the ARTP method to combine these gene-level P-values into a test statistic for the diseasepathway association. The difference between the gene-based and SNP-based approaches stems from whether the genes or the SNP are treated equally in defining the final pathway-level test statistics.

The ARTP method [Dudbridge and Koeleman 2004; Hoh, et al. 2001] is a general P-value combining approach that is designed to overcome the following major drawback of the standard RTP method. To use the RTP method, the user typically specifies a rank truncation point *K* and combines the top *K* most significant P-values as the summary statistic. The selection of the rank truncation point can be quite arbitrary, especially when there are a large number of P-values to be combined. Choosing a truncation point too large or too small can have a detrimental effect on the power. The ARTP method, on the other hand, optimizes the selection of the truncation point among a set of candidates. The final test statistic is the minimum empirical P-value observed at various candidate truncation points. The idea of using the minimum P-value as the test statistic is commonly used in genetic studies [Chen, et al. 2006; Dudbridge and Koeleman 2004; Gonzalez, et al. 2008; Hoh, et al. 2001; Li, et al. 2008; Zheng and Chen 2005].

A challenge for the proposed gene-based pathway analysis is that it potentially requires a multiple-layer of re-sampling procedure to evaluate the significance level of the pathway-level test statistic. First, depending on the choice of the test statistic, an inner level of permutation may be required to obtain the gene-level summary of association. Then, a second level of permutation may be needed to obtain the P-value associated with the pathway-level RTP statistic for a given truncation point. Finally, a third level of permutation may be needed to evaluate the significance level of the ARTP statistic after accounting for multiple testing over different truncation points. Clearly, such multi-level permutation procedure can quickly become computationally infeasible as the number of genes considered in a pathway becomes large (say, over a few hundreds).

To overcome the computational obstacles mentioned above, we propose an efficient algorithm to evaluate the significance level of the gene-based pathway analysis. The algorithm adopts the idea of Ge et al. [2003] that uses a single level of permutation iterations to achieve the goal of the multiple-level permutation procedure.

We describe a novel approach to simulate pathway-based case-control association study by mimicking gene-structure and LD patterns observed from empirical data and use this procedure to evaluate performance of the proposed methodology in realistic settings. As an application, we apply the proposed methods to study the association between the nicotinic receptors pathway and cigarette smoking behaviors based on a joint analysis of two existing GWAS [Hunter, et al. 2007; Yeager, et al. 2007].

Methods

Adaptive rank truncated product (ARTP) method

Suppose we conduct L tests on a global null hypothesis $\mathbf{H}_0$, and let the resulting P-values be $p_1,\ldots, p_L$. For example, if we want to test the null hypothesis that a pathway consisting of L SNPs is not associated with a disease outcome, we can perform tests on individual SNPs, e.g., using the standard Cochran-Armitage trend test [Agresti 2002] within the considered pathway. Denote the ordered statistics of those P-values by $p_{(1)} \leq \ldots \leq p_{(L)}$, with $p_{(l)}$ being the lth smallest P-value. To test for a global null hypothesis, the rank truncated product (RTP) statistic [Dudbridge and Koeleman 2003] has been proposed with the form,

$$W(K) = \prod_{i=1}^{K} p_{(i)},$$

with K, $1 \leq K \leq L$, being a predetermined integer (the truncation point). In words, $W(K)$ simply is the product of the K smallest P-values. When all tests are independent, the P-value associated with $W(K)$ can be obtained analytically. If they are correlated, e.g., due to LD among the SNPs with a gene, a permutation procedure is generally needed to obtain the significance level of $W(K)$.

To use this rank truncated product statistic, one must choose the truncation point K *a priori*. When the number of individual tests is large (say over 100), it is difficult to make a sensible choice of K. Instead, we can optimize the association evidence obtained on each of J candidate truncation points $K_1 \leq \ldots \leq K_J$. More specifically, let $\hat{S}(Kj)$ be the estimated P-value for $W(Kj)$, $W(K_j)$,$1 \leq j \leq J$, the following statistic based on minimum P-value can be defined [Dudbridge and Koeleman 2004; Hoh, et al. 2001].

$$MinP = \min_{1 \leq j \leq J} \widehat{S}\left(K_j\right).$$

In the following discussion we call this the adaptive rank truncation product (ARTP) method. To limit the effect of multiple comparisons we recommend using approximately $J = 10$ candidate truncation points. Some examples of choosing candidate truncation points are given in the following sections. For a single truncation point (i.e., $J = 1$), the ARTP method becomes the standard RTP method.

To get the adjusted P-value for *MinP*, one generally needs a two-level permutation procedure [Hoh, et al. 2001; Westfall and Young 1993] with the inner level for estimating $\widehat{S}\left(K_j\right)$ and the outer level for the adjustment needed to account for multiple testing over different truncation

points. This type of permutation procedure, however, can become computationally infeasible if the number of tests L is relatively large. We propose to use a single layer of permutation for determining the significance level of the ARTP statistic by borrowing techniques originally introduced for gene-expression data analysis [Ge, et al. 2003]. We first obtain P-values for each test on the null hypothesis based on the observed data, and denote them as $p_1^{(0)}, \cdots, p_L^{(0)}$. Then we use an appropriate permutation (or resampling) procedure to generate B datasets under the null hypothesis $\mathbf{H}_0$. Based on the b^{th} permuted dataset, $1 \le b \le B$, we perform the L individual tests and denote the resultant P-values as $p_1^{(b)}, \cdots, p_L^{(b)}$. The permutation procedure used for generating null datasets is application-specific. We will provide some examples in the following discussion. Based on those obtained P-values, we apply the following steps to obtain the adjusted P-value for *MinP*.

The *MinP* Algorithm:

1. Based on $p_1^{(b)}, \cdots, p_L^{(b)}$, $0 \le b \le B$, calculate the rank truncated product statistics for each candidate truncation point, and denote them as $W_j^{(b)} = \prod_{i=1}^{K_j} p_{(i)}^{(b)}$, $1 \le j$.
2. Based on $W_j^{(b)}$, $1 \le j \le J$, $0 \le b \le B$, use Ge's algorithm (see below for detail) to obtain the estimated P-value $\widehat{S}_j^{(b)}$ corresponding to $W_j^{(b)}$.
3. Let $MinP^{(b)} = \min_{1 \le j \le J} \widehat{S}_j^{(b)}$, $0 \le b \le B$. The adjusted P-value for the adaptive rank truncated product statistic $MinP^{(0)}$ is estimated as $\frac{\sum_{b=0}^{B} I\left(MinP^{(b)} \le MinP^{(0)}\right)}{B+1}$.

In Step 2, we need to calculate the estimated P-value $\widehat{S}_j^{(b)}$ for $W_j^{(b)}$ from the observed data (b = 0) as well as the b^{th} ($\ge$1) generated dataset. The standard approach would apply another level of permutations specifically based on the b^{th} ($\ge$1) permuted dataset. To avoid this, following Ge et al. [Ge, et al. 2003] we use the set $\left\{W_j^{(b)}, b=0, \cdots, B\right\}$ itself to form a common reference distribution for the evaluation of the significance level of each $W_j^{(b)}$. More specifically, we estimate the P-value for $W_j^{(b)}$ as $S_j^{(b)} = \frac{\sum_{b^*=0}^{B} I\left(W_j^{(b^*)} \le W_j^{(b)}\right)}{B+1}$. Note that the algorithm combines statistics derived from the observed dataset as well as those from permuted datasets to form the reference distribution for the evaluation of P-values, following the suggestion by [Becker and Knapp 2004]. In Step 3, we estimate the adjusted P-value for the observed ARTP statistic $MinP^{(0)}$.

In a calculation similar to that in Step 2, we can simultaneously estimate the adjusted P-value for $MinP^{(b)}$, the ARTP statistic from the b^{th} generated dataset, $1 \le b \le B$, by $\frac{\sum_{b^*=0}^{B} I\left(\text{MinP}^{(b^*)} \le MinP^{(b)}\right)}{B+1}$.

Gene-based pathway analysis

Motivation for gene-based pathway analysis—We are interested in evaluating the association between a phenotype and a pathway that consists of a set of candidate genes. There are usually multiple SNPs within each candidate gene. The null hypothesis is that there is no association between the phenotype and any of the SNPs included in the genetic pathway. Both SNP-based and gene-based tests could be applied to test this null hypothesis. The SNP-based strategy involves first performing an association test for each individual SNP and then

combining the SNP-level evidence of associations through a P-value combining method, such as the ARTP or RTP method described earlier. The gene-based strategy is first to conduct a gene-level association test on the association between each gene and the phenotype, and then to combine gene-level evidence using a P-value combining method. The gene-level association test itself could involve first testing for single-SNP association within the gene and then combining the SNP-level associations within the gene. Alternatively, it could also involve a multi-locus test that analyzes all the SNPs within a gene simultaneously [Fallin, et al. 2001; Fan and Knapp 2003; Schaid, et al. 2002; Tzeng, et al. 2003; Xiong, et al. 2002].

The gene-based approach treats the association signal from each gene equally, while the SNP-based approach values evidence from each SNP equally. The following two extreme examples illustrate the potential advantages and disadvantages of the two approaches. Suppose a pathway consists of just two genes, one a large disease-unrelated gene with 100 SNPs, the other a small disease-associated gene with only one SNP. In the gene-based approach, 1 out of 2 testing units (individual genes) contains the association signal, while in the SNP-based approach, only one out of 101 testing units (individual SNPs) conveys the association signal. Thus, we expect the gene-based approach to be more powerful, since it limits the influence of many null SNPs in the large gene. On the other hand, suppose a pathway consists of one large disease related gene containing multiple susceptibility SNPs (say 10 out of 100 SNPs), and 100 small disease unrelated genes with just 1 SNP per gene. In the gene-based approach, only 1 out of 101 testing units (individual genes) contains the association evidence. In the SNP-based approach, however, 10 out of 200 testing units (SNPs) show evidence of association. Thus, we expect in this case that the SNP-based approach could be more powerful than the gene-based approach. In real applications, the relative merit of the two approaches depends on the signal-to-noise ratio in all testing units, as well as on the LD patterns in each gene. In the simulation study, we compare these two approaches under more realistic scenarios.

Gene-based pathway analysis using the ARTP method—We propose to use the ARTP method described above to combine gene-level P-values across all candidate genes included in a biological pathway. The gene-level P-value corresponds to a chosen statistic for the candidate gene association test and represents the significance level of the association between a gene and the outcome. Depending on the test statistic, the gene-level P-value may or may not require a separate permutation procedure. When the gene-level P-value can be evaluated through a known asymptotic distribution, it is straightforward to apply the ARTP method to combine gene-level P-values for the pathway analysis. When the chosen gene-level test requires a permutation procedure for the evaluation of the gene-level P-value, we again can adopt the idea of Ge et al. [Ge, et al. 2003] to avoid the computationally intensive multi-level permutation procedure. Below we provide a summary of the steps for performing the gene-based pathway analysis if the ARTP method is used both to derive the gene-level summary and to combine gene-level P-values across all genes. The algorithm can be modified slightly for other gene-level summaries requiring permutation.

Suppose the pathway consists of L genes, with the l^{th} consisting of n_l SNPs, $1 \le l \le L$. Let $p_{l,i}^{(0)}$ be the P-value for the association test on the i^{th} SNPs of the l^{th} gene based on the observed dataset. We then perform an appropriate permutation procedure to generate B datasets under the null hypothesis, and let $p_{l,i}^{(b)}$ be the P-value for the test on the i^{th} SNPs of the l^{th} gene based on the b^{th} generated dataset, $1 \le b \le B$. Here we assume the SNP-level P-value $p_{l,i}^{(b)}$ can be obtained without a permutation procedure. Next, we apply the ARTP (with a predetermined set of candidate truncation points, which could be varied from gene to gene) to combine SNP-level evidence of association within a gene. More specifically, for the l^{th} gene, we apply the *MinP* algorithm given earlier on $p_{l,i}^{(b)}$, $1 \le i \le n_l$, $0 \le b \le B$, to obtain $p_l^{*(0)}$, the gene-level P-

value for the observed data, and $p_l^{*(b)}$, the gene-level P-value for the b^{th} permuted dataset. Finally, as a pathway analysis we apply the ARTP statistic (with pre-selected candidate truncation points) to combine the gene-level P-values for the observed and permuted "null" data sets and use the *MinP* algorithm once again to obtain the adjusted P-value for the pathway-level ARTP statistic. As before, the key here is that the P-values for the test statistics derived from the observed dataset as well as permuted datasets are obtained by using a common set of referent statistics, thus avoiding computationally expensive multi-layer permutation algorithms. In this procedure, the same set of generated "null" data sets are used in the *MinP* algorithm for both the gene-level and pathway-level summaries

Since the RTP is a special case of the ARTP with just one truncation point, the above algorithm can be applied to the gene-based pathway analysis when the RTP is used for the gene-level summary.

Other pathway analysis approaches and their definitions

In the simulation studies and application, we consider various versions of pathway analysis methods. We use X-Y to represent a gene-based pathway analysis that uses X (the RTP or ARTP) as the gene-level summary statistic combining SNP-level P-value and Y (the RTP or ARTP) as the pathway-level summary statistic combining gene-level P-values. When the ARTP is used for the gene-level summary, we use a set of 5 candidate truncation points with $K_j = J$ ma x{1, [n/2 0]} $j = 1, \ldots, 5$, where n denotes the number of SNPs in the gene, and $[n / 20]$ represents the largest integer that is less or equal to $[n / 20]$. Thus for a gene with less than 20 SNPs, we inspect its top 1, 2, …, 5 SNPs, otherwise, we use 5 truncation points, with each at every $[n / 20]$ SNPs. When the RTP method is used for the gene-level summary, we look at just one truncation point with $K_1 = 1$.

We use ARTP (or RTP) to denote the SNP-based pathway analysis that applies the ARTP (or RTP) statistic to combine SNP-level P-values among SNPs in the pathway, ignoring the underlying gene structure.

Simulation Studies

Simulation design

Following, we first describe a novel way of simulating pathway-based case-control association study imitating gene-structure and LD patterns from real data. Suppose a pathway consists of L genes, with the l^{th} gene having n_l SNPs, $1 \le l \le L$, let $G_l = (g_{l,1}, \cdots, g_{l,n_l})$ be the random variable vector representing the joint genotype for the l^{th} gene, and let U_l be the observed value for G_l in a subject. We assume that in the source population, all genes in the considered pathway are independent. The assumption is not required for the method, but is used only for the ease of simulation. Instead of choosing arbitrary values for $\Pr(G_l)$, $1 \le l \le L$, we used genotypes generated from the prostate cancer GWAS in the Cancer Genetic Markers of Susceptibility (CGEMS) project for the specification of $\Pr(G_l)$, $1 \le l \le L$. In the prostate cancer GWAS, approximately 523,000 SNPs on the Illumina platform were measured on 2,329 men from the Prostate, Lung, Colon and Ovarian (PLCO) Trial [Yeager, et al. 2007].

For the disease model, we assume the first M genes are associated with the disease, with each having a single functional SNP that directly affects the disease risk. The risk model we adopt in our simulation is

$$\text{logit}\left[\Pr\left(Y=1|g_1^*, \cdots, g_M^*\right)\right] = \alpha + \sum_{l=1}^{M} \beta_l g_1^*, \quad (1)$$

where g_l^* is the genotype (coded as 0, 1 or 2 according to counts of minor alleles) at the functional SNP in the l^{th} gene, $1 \le l \le M$. Under this disease model and the assumption of a rare disease as well as the independence of genotypes at different genes, the M functional SNPs have multiplicative effect on the risk of the disease and one can generate the genotype data for each gene independent of the others, for both the controls and the cases (see Appendix). This greatly simplifies the simulation procedure.

We simulate the joint genotype on a disease-unrelated gene (say the l^{th} gene with $l > M$) for both cases and controls according to $\Pr(G_l)$ observed in the real study (the CGEMS prostate cancer GWAS). In particular, we randomly draw a subject from the real study, and assign his joint genotype for the given gene to the simulated subject. The genotype data for a disease related gene can be generated in the same manner as above for simulated controls. For cases, however, the joint genotype distribution of a disease related gene will be altered from $\Pr(G_l)$ according to the specified disease-risk model. In particular, for a case, we first notice that the joint genotype probability $\Pr(G_l = U_l | Y = 1)$ can be expressed as $\Pr\left(g_l^* = u_l^* | Y = 1\right) \Pr\left(G_l = U_l | g_l^* = u_l^*\right)$, with u_l^* being the element corresponding to the functional SNP in U_l, under the assumption that given g_l^* the genotype for the remaining SNPs in a gene are independent of the outcome of interest. Under the disease model (1) and the rare disease assumption, we further have that

$$\Pr\left(g_l^* = u_l^* | Y = 1\right) = \frac{\exp\left(\beta_l u_l^*\right) \Pr\left(g_l^* = u_l^*\right)}{\Pr\left(g_l^* = 0\right) + \exp\left(\beta_l\right) \Pr\left(g_l^* = 1\right) + \exp\left(2\beta_l\right) \Pr\left(g_l^* = 2\right)}. \quad (2)$$

Thus, we can simulate the joint genotype on a disease-associated gene for a case by the following three steps: 1. Generate the genotype (denoted as u_l^*) at its functional locus according to (2); 2. Randomly select a subject among subjects having genotype u_l^* at the functional SNP in the real study; and 3. Assign the joint genotype on this gene observed at the selected subject to the simulated case.

We considered two hypothetical pathways, each defined as a set of unlinked candidate genes chosen from a total list of 359 candidate genes potentially involved in smoking behaviors (Caporaso et al., unpublished manuscript). Although some candidate genes could actually be linked, we still assume they are unlinked when assigning genotypes in our simulation. First, we considered a pathway consisting of 110 genes, with each gene consisting of roughly the same number of SNPs (10-20). There are a total of 1,517 SNPs in this pathway. We choose M, the number of disease-related genes, to be 0, 5, 15, 30, or 45; and assign a SNP near the center of each disease-associated gene as the functional SNP. For each situation with $M > 0$, we further consider two scenarios, one with genotypes measured, the other with genotypes unmeasured at all disease risk SNPs. We choose a common value for all the β_l, $1 \le l \le M$, in the model (1) for a given M in such a way that the power of the ARTP-ARTP falls into an appropriate range.

The second hypothetical pathway consists of 12 genes containing a total of 516 SNPs, with the number of SNPs per gene varying considerably. Among the 12 genes, four are small-sized (in term of the number of SNPs included) with 8 SNPs in each gene, four are medium-sized, with 33-37 SNPs in each gene; and the remaining are large-sized, with 75-96 SNPs within each gene. We let M, the number of disease related genes, be 0 for type I error evaluation, and 4 for the power evaluation under the following three disease-gene assignments: 1. all 4 small-sized genes are disease-related; 2. all 4 mediumsized genes are disease-related; and 3. all 4 large-

sized genes are disease-related. We let the $\beta_l = \log(1.25)l$, = 1,…,4, for each of the 4 functional SNPs in the model (1).

To evaluate the type I error of the considered tests under the proposed simulation design, we simulated 2,000 datasets, with each consisting of 1,000 cases and 1,000 controls, where the case-control status could indicate, for example, any binary smoking related trait of interest, such as current vs. former smoking status. For the power evaluation, we simulated 500 datasets, with each consisting of 1,000 cases and 1,000 controls.

Results

Type I error—For each simulated dataset, we applied two types of gene-based pathway analyses, that is, RTP-ARTP and ARTP-ARTP, as well as the SNP-based pathway analysis using ARTP with various sets of candidate truncation points. The configuration detail for each method is given in Table 1. Empirical type I error rates for the significance level of 0.05 are summarized in Table 2. It can be seen from Table 2 that all tests have their empirical type I error rate close to the nominal level based on 2,000 replications.

Power comparison between the gene-based and SNP-based approaches for pathway analysis—We first focused on the pathway that consists of 12 genes of varying size. The power comparison under each of the three considered simulation designs was based on 500 simulated datasets, with each consisting of 1,000 cases and 1,000 controls. Results are summarized in Table 3. When the disease-associated genes happen to have more SNPs than the disease-unrelated genes do, the SNP-based approach (ARTP-I, -II or -III, and RTP(q)) appears to be slightly more powerful than either of the two gene-based approaches. When the disease-associated genes are medium-sized (compared with other genes in the pathway in terms of the number of SNPs contained), the power of the SNP-based and gene-based approaches are comparable, with the power of ARTP-ARTP slightly higher than the SNP-based approaches. When the disease-associated genes are relatively small, both gene-based approaches are much more powerful than the SNP-based approaches. The percentage increase in power can be as much as 80%. Thus, between the two gene-based approaches, ARTP-ARTP is more powerful than RTPARTP, especially when functional SNPs are not measured.

We also conducted simulations under the pathway that one consists of 110 of approximately the same size. We found that the gene-based and SNP-based approaches have comparable power under various values for M, the number of disease-related genes (results not shown). This observation is consistent with that observed in the 12 gene pathway when the disease-related genes are medium-sized. Also, ARTP-ARTP appears to be more powerful than RTP-ARTP (results not shown).

In summary, neither the gene-based approach nor the SNP-based approach dominates the other in all considered cases. However, the gene-based approach ARTPARTP appears to be more robust than SNP-based approaches in the sense that it has either optimal or close to optimal power in all different scenarios considered. In contrast, the SNP-based analysis approach can suffer major loss of power compared to the gene-based method for studying pathways consisting of genes with highly variable sizes

Power comparison between the adaptive and fixed truncation-point approaches for pathway analysis—In the gene-based pathway analysis, once we have summarized the gene-level effect by the ARTP, we can combine evidence across all genes using either the adaptive truncation point strategy (ARTP-ARTP), or the fixed truncation point strategy (ARTP-RTP(q), with q being the fixed truncation point). We conducted simulation studies to compare these two approaches. Simulation results for the pathway consisting of 110 genes under various values of M are summarized in Table 4. Among all 9 considered methods, ARTP-

ARTP clearly has the overall best performance across the various scenarios considered. There is no single truncation point *q* that gives optimal or close-to-optimal power for the RTP procedure in all the different scenarios considered. In each setting, however, the power of the ARTP always remains close to that for the optimal RTP for that setting. If RTP, instead of ARTP, was used to summarize the gene-level effect, by comparing RTP-ARTP with RTP-RTP (*q*) we reach the same conclusion, that the adaptive approach is superior to the fixed approach for summarizing association evidence across all genes (results not shown). Simulation studies conducted for the pathway consisting of 12 genes under all three considered disease-associated gene assignments (disease-associated genes are large-, medium- or small-sized) also support this conclusion (results not shown).

Application: the association between the nicotinic receptors pathway and smoking behaviors

Caporaso et al. (unpublished manuscript) conducted a GWAS of smoking phenotypes using existing data for 2,329 men from the Prostate, Lung, Colon and Ovarian (PLCO) Trial, and 2,282 women from the Nurses' Health Study (NHS), all of whom were previously genotyped in the Cancer Genetic Markers of Susceptibility (CGEMS) project. They tested for association between each SNP and each of several measures of smoking behavior, including cigarettes per day (CPD), and duration of smoking (SMKDU) among current or former smokers, with the analysis adjusted for age, cohabitation/marital status, education, study site, and the top principal components of the population structure.[Price, et al. 2006] Besides performing a whole genome scan, the study also selected, *a priori*, 359 candidate genes potentially involved in smoking behaviors, for more in depth analysis. Most of the candidate genes were nominated by the Candidate Gene Committee of NICSNP (http://zork.wustl.edu/nida/Results/cand_genes.htm). Using a slightly modified version of the groups developed by the NICSNP, the study defined 30 pathways with 3-34 genes per pathway, and performed pathway analysis using the standard RTP method.

To demonstrate the application of the proposed ARTP method, we focused on the defined nicotinic receptors pathway that consists of 16 candidate genes with a total of 159 SNPs (Table 5) and studied its association with smoking behavior phenotypes CPD and SMKDU. For a given SNP, we tested for its association with a phenotype (log-transformed CPD or SMKDU) using linear regression, adjusting for age, cohabitation/marital status, education, site, and the top principal components of the population structure. SNP genotypes were coded as counts of minor alleles, leading to a 1 degree of freedom trend test. The same test was conducted separately for PLCO and NHS. We combined evidence for association across PLCO and NHS using a weighted Z test, and obtained the corresponding P-value for each of 159 selected SNPs in the nicotinic receptors pathway. Results from the single-marker analysis were given in Caporaso et al. [2009]. For the pathway analysis, we used the residual permutation procedure to generate the null dataset adjusting for covariates [Anderson and Legendre 1999]. More specifically, we first fit the reduced linear regression model using variables other than the SNP genotype, then permuted the residuals and add permuted residuals to the fitted values to generate the new outcomes. The null dataset consisted of the new outcomes as well as the original genetic and non-genetic predictive variables. The rationale for using such a permutation procedure was to maintain the correlation between the adjusted covariates (e.g., principal components) and the genotypes.

We applied both the SNP-level and gene-level ARTP pathway analysis with 20,000 permutations. For the SNP-based pathway analysis, we focused only on the ARTP method, as it is difficult to choose a fixed truncation point for the RTP method. For the gene-based pathway analysis RTP-ARTP, we first used RTP to summarize the gene-level evidence by focusing on only the smallest P-value observed among SNPs within each gene and then used ARTP to

summarize the pathway effect by optimizing association evidence over inspections at the top 1, 2,…, and 10 ranked genes. For ARTP-ARTP, we first used ARTP to summarize the gene-level evidence by using candidate truncation points at the top 1, 2,, 3, 4 and 5 SNPs and then used the same ARTP procedure for the pathway effect summary as the one used in RTP-ARTP. For the SNP-based pathway analysis, to be comparable with the gene-based analysis, we considered 3 different ARTP approaches, each with 10 truncation points, including ARTP-I with a truncation point at every SNP, ARTP-II with a truncation point at every 5 SNPs, and ARTP-III with a truncation point at every 10 SNPs.

Results are summarized in Table 6. There is evidence suggesting an association between the nicotinic receptors pathway and the outcome CPD with all tests having P-values less than 0.05. The association between the considered pathway and the outcome SMKDU is also borderline significant. We can also notice from the Table 6 that gene-based analyses consistently yielded smaller P-values than those from the SNP-based analyses.

Discussion

Benefiting from recent advances in high-throughput genotyping technology, researchers are now able to evaluate hundreds of thousands of SNPs throughout the genome in search for disease-susceptible loci. This agnostic single-SNP testing approach enables one to discover disease-associated loci in regions of genome where scientists have little or no prior knowledge about the biologic function (e.g., the 8q24 “gene-desert” region that has been recently associated with the prostate cancer [Yeager, et al. 2007]). A complementary and potentially rewarding strategy for genetic association studies is to study jointly the association between the trait and multiple genetic variants within a pathway defined according to the current biological knowledge.

In this report, we proposed a class of pathway analysis approaches based on the adaptive rank truncated product (ARTP) statistic and an associated computationally efficient permutation algorithm for evaluating the significance of the test-statistics. Through simulation studies, we compared gene-based and SNP-based strategies for the pathway analysis and found that the former approach have more robust performance. In particular, we found that when a pathway consists of genes of highly variable sizes, the gene-based method can have a major power advantage over its SNP-based counterpart if the causal variants reside in the relatively smaller sized genes. In contrast, there was no setting in which the SNP-based approach was clearly more superior among the two methods we compared. Furthermore, we found an adaptive approach for choosing the truncation point for rank-truncated-product statistics can improve the power of the method compared to fixing the truncation point to a pre-defined value. These observations were further reinforced in an application of considered methods for the study of the association between smoking behavior and nicotinic receptor pathway.

The ARTP method provides an efficient and flexible way to accumulate association evidence across individual genes within a pathway. In this report, we have obtained gene-level summary of association by combining results from the single-SNP test-statistics, using a RTP or an ARTP method, within a gene. Alternatively, one can obtain gene-level summary by constructing a multi-locus test for association that involves simultaneous analysis of all the SNPs within a gene. A variety of such powerful methods have been recently become available in the literature [Gauderman, et al. 2007; Kwee, et al. 2008; Schaid, et al. 2005; Yu, et al. 2004; Yu, et al. 2005; Zaykin, et al. 2006]. The proposed ARTP procedure could be easily adapted based on these alternative multi-locus test-statistics. The method is computational feasible even if the evaluation of the gene-level P-value associated with the chosen multi-locus test requires a permutation procedure, since only a single-level permutation procedure is needed for

evaluating the significance of the final test-statistics. This offers great flexibility in incorporating a wide range of gene-level or SNP-level summary statistics.

Dudbridge and Koeleman [2004] proposed a computationally efficient permutation approach to evaluate the significance level for the ARTP statistic based on the extreme-value distribution theory. But their approach is appropriate only when the number of testing units (e.g., the genes in the gene-based pathway analysis) is much larger than considered truncation thresholds. Thus their method is most suitable for GWAS where we expect only a handful of true disease-association SNPs among over 100,000 testing SNPs, but not for pathway analysis where the number of genes could range between ten to a few hundreds. Also, a multi-level permutation procedure cannot be avoided by using their approach when the evaluation of the gene-level P-value itself requires a separate permutation procedure.

Besides the RTP, there are other types of P-value combination approaches that use a fixed truncation point, such as the one proposed by Zaykin et al. [2002] that combines P-values less than a given threshold. It is possible to develop an adaptive version of this method using the algorithm described in this report.

Recently, gene set enrichment analysis (GSEA) algorithm [Subramanian, et al. 2005] has been proposed for the identification of disease related pathways by measuring the overrepresentation of disease-gene associations within a given pathway compared to a list of reference genes [Wang, et al. 2007]. The underlying null hypothesis is that the set of genes in a given pathway has no enrichment of association signals compared to the rest. In contrast, in this report, we focus on testing for the effect of a specific pathway without reference to any larger gene list. The underlying global null hypothesis is that there is no association of the disease with any of the genes in the given pathway. We believe that for GWAS, where the vast majorities of the reference genes are likely to be unrelated to a particular trait, the “global” vs. the “enrichment” null hypotheses are approximately the same and both types of approaches could be valuable for testing and prioritizing candidate disease susceptibility pathways. Many of the statistical and computation issue regarding how to combine evidence of association from SNPs to genes and then to pathways are similar between the two approaches. Thus, some of the tools we utilized, including the efficient permutation algorithm, could be useful for GSEA type analysis as well.

Several areas of research remain open. For example, once the association between a pathway and an outcome has been established, methods are needed for identifying the specific subset of the genes and the SNPs within the genes that are actually responsible for the association. This task can get particularly challenging partially due to LD between physically nearby SNPs and genes.

The proposed P-value combining methods gain efficiency by accumulating marginal association signal across individual testing units. Although a particular disease model was chosen for the simulation studies, we think the general conclusion still hold as long as there is association evidence from individual testing units. This approach, however, may not be very powerful in situations when there is none or very weak marginal effects from the individual genes, but there is strong epistatic interactions among the genes. In principle, the proposed ARTP procedure can be also adapted to account for gene-gene or/and SNP-SNP interactions within a pathway. For example, one can consider ARTP statistics by accumulating evidence of associations over joint analysis of pairs of genes within a pathway. A variety of advanced methods can be used to allow for epistatic interactions in such joint analysis [Chapman and Clayton 2007; Chatterjee, et al. 2006; Chen, et al. 2007; Ritchie, et al. 2001; Ruczinski, et al. 2003; Zhao, et al. 2006].

In summary, the proposed gene-based ARTP procedure, given its power, flexibility and computational efficiency, is a promising approach for pathway-based association analysis. We believe that in future this efficient single-level permutation algorithm will allow the method to adapt itself to incorporate more complex information, such as epistatic interactions or biologic knowledge about gene networks, into pathway analysis without increasing the associated computational burden dramatically.

Acknowledgments

This research utilized the high-performance computational capabilities of the Biowulf PC/Linux cluster at the National Institutes of Health, Bethesda, Maryland, USA (http://biowulf.nih.gov). The work of K Yu, Q Li, RM Pfeiffer, PS Rosenberg, N Caporaso, and N Chatterjee were supported in part by the Intramural Program of the NIH and the National Cancer Institute. The work of Q Li was also partially supported by the National Science Foundation of China, No. 10371126. AW Bergen is supported by U01 DA020830.

Appendix: Genotype independence in cases and controls

Here we provide more technique details justifying the genotype independence in cases, as well as in controls. Following the argument of Gail et al.,[Gail, et al. 2008] we first derive a general result under the following two assumptions: I) in the source population all genes are independent, i.e., $\Pr(G_1=U_l,\cdots,G_L=U_L)=\prod_{l=1}^{L}\Pr(G_l=U_l)$ and II) the risk for the disease satisfies the condition that $\Pr(Y=1|G_1=U_1,\cdots,G_L=U_L)=c_0\prod_{l=1}^{L}f_l(U_l)$, where Y represents the outcome, and c_0 is a constant value.

Based on these two assumptions, we have

$$\Pr(Y=1)=c_0\sum_{V_1,\cdots,V_L}\prod_{l=1}^{L}f_1(V_l)\Pr(G_l=V_l)=c_0\prod_{l=1}^{L}\left(\sum_{V_l}f_l(V_l)\Pr(G_l=V_l)\right), \tag{A1}$$

$$\Pr(G_1=U_1,\cdots,G_L=U_L,Y=1)=c_0\prod_{l=1}^{L}f_l(U_l)\Pr(G_l=U_l). \tag{A2}$$

From (A2), it can be shown that for $1\le l\le L$,

$$\Pr(G_l=U_l,Y=1)=c_l f_l(U_l)\Pr(G_l=U_l), \tag{A3}$$

where $c_l=c_0\prod_{k\neq l}^{L}\left(\sum_{V_k}f_k(V_k)\Pr(G_k=V_k)\right)$. Using (A3), we can expressas $\Pr(G_l=U_l|Y=1)$ as

$$\Pr(G_l=U_l|Y=1)=\frac{f_l(U_l)\Pr(G_l=U_l)}{\sum_{V_l}f_l(V_l)\Pr(G_l=V_l)}. \tag{A4}$$

Finally, according to A1, A2 and A4, we have

$$\begin{aligned}\Pr(G_1=U_1,\cdots,G_L=U_L|Y=1) &= \frac{\Pr(G_1=U_1,\cdots,G_L=U_L,Y=1)}{\Pr(Y=1)}\\ &= \prod_{l=1}^{L}\frac{f_l(U_l)\Pr(G_l=U_l)}{\sum_{V_l} f_l(V_l)\Pr(G_l=U_l)}\\ &= \prod_{l=1}^{L}\Pr(G_l=U_l|Y=1).\end{aligned}$$

Thus, under the assumption I and II, we have shown

$$\Pr(G_1=U_1,\cdots,G_L=U_L|Y=1)=\prod_{l=1}^{L}\Pr(G_l=U_l|Y=1). \quad (A5)$$

.

For a rare disease with its risk model given by (1) in the Text, we have $\Pr\left(Y=1|g_1^*,\cdots,g_M^*\right)\approx\exp\left(\alpha+\sum_{m=1}^{M}\beta g_m^*\right)$. Thus the disease model we adopted satisfies the assumption II. As a result, we can generate genotypes for cases according to (A4) and (A5) Also, under the assumption of a rare disease, the controls can be thought as a random sample from the source population, i.e.,

$$\Pr(G_1=U_1,\cdots,G_L=U_L|Y=0)\approx\Pr(G_1=U_1,\cdots,G_L=U_L)=\prod_{l=1}^{L}\Pr(G_l=U_l).$$

Reference

Agresti, A. Categorical data analysis. Wiley: 2002.

Anderson MJ, Legendre P. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical computation and simulation 1999;62(3):271–303.

Becker T, Knapp M. A powerful strategy to account for multiple testing in the context of haplotype analysis. Am J Hum Genet 2004;75(4):561–70. [PubMed: 15290652]

Caporaso N, Gu F, Chatterjee N, Jin S, Yu K, Yeager M, Hunter DJ, Jacobs K, Landi MT, Chanock S, Ziegler R, Hankinson S, Chen J, Kraft P, Bergen AW. Genome-wide and candidate gene association scans of cigarette smoking behaviors. PLoS ONE. in press

Chapman J, Clayton D. Detecting association using epistatic information. Genet Epidemiol 2007;31(8): 894–909. [PubMed: 17654599]

Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet 2006;79 (6):1002–16. [PubMed: 17186459]

Chen BE, Sakoda LC, Hsing AW, Rosenberg PS. Resampling-based multiple hypothesis testing procedures for genetic case-control association studies. Genet Epidemiol 2006;30(6):495–507. [PubMed: 16755536]

Chen J, Yu K, Hsing A, Therneau TM. A partially linear tree-based regression model for assessing complex joint gene-gene and gene-environment effects. Genet Epidemiol 2007;31(3):238–51. [PubMed: 17266115]

Consortium WTCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447(7145):661–78. [PubMed: 17554300]

Dudbridge F, Koeleman BP. Rank truncated product of P-values, with application to genomewide association scans. Genet Epidemiol 2003;25(4):360–6. [PubMed: 14639705]

Dudbridge F, Koeleman BP. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am J Hum Genet 2004;75 (3):424–35. [PubMed: 15266393]

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 2001;11(1):143–51. [PubMed: 11156623]

Fan R, Knapp M. Genome association studies of complex diseases by case-control designs. Am J Hum Genet 2003;72(4):850–68. [PubMed: 12647259]

Fisher, RA. Statistical methods for research workers. London Oliver and Boyd; 1932.

Gail MH, Pfeiffer RM, Wheeler W, Pee D. Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. Biostatistics 2008;9(2):201–15. [PubMed: 17873152]

Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. Genet Epidemiol 2007;31(5):383–95. [PubMed: 17410554]

Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. Test 2003;12 (1):1–44.

Gonzalez JR, Carrasco JL, Dudbridge F, Armengol L, Estivill X, Moreno V. Maximizing association statistics over genetic models. Genet Epidemiol 2008;32(3):246–54. [PubMed: 18228557]

Hoh J, Wille A, Ott J. Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res 2001;11(12):2115–9. [PubMed: 11731502]

Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007;39(7):870–4. [PubMed: 17529973]

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet 2008;82(2):386–97. [PubMed: 18252219]

Li Q, Zheng G, Li Z, Yu K. Efficient approximation of P-value of the maximum of correlated tests, with applications to genome-wide association studies. Ann Hum Genet 2008;72(Pt 3):397–406. [PubMed: 18318785]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38(8):904–9. [PubMed: 16862161]

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 2001;69(1):138–47. [PubMed: 11404819]

Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. Journal of computational and graphical statistics 2003;12(3):475–511.

Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN. Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet 2005;76(5):780–93. [PubMed: 15786018]

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002;70(2):425–34. [PubMed: 11791212]

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102(43):15545–50. [PubMed: 16199517]

Tzeng JY, Devlin B, Wasserman L, Roeder K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet 2003;72(4):891–902. [PubMed: 12610778]

Wang K, Li M, Bucan M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. Am J Hum Genet 2007;81(6)

Westfall, PH.; Young, BS. Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment,. Willey; New York: 1993.

Xiong M, Zhao J, Boerwinkle E. Generalized T2 test for genome association studies. Am J Hum Genet 2002;70(5):1257–68. [PubMed: 11923914]

Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007;39(5):645–9. [PubMed: 17401363]

Yu K, Gu CC, Province M, Xiong CJ, Rao DC. Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes. Genet Epidemiol 2004;27(3):182–91. [PubMed: 15389925]

Yu K, Xu J, Rao DC, Province M. Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. Ann Hum Genet 2005;69(Pt 5):577–89. [PubMed: 16138916]

Zaykin DV, Meng Z, Ehm MG. Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. Am J Hum Genet 2006;78(5):737–46. [PubMed: 16642430]

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining P-values. Genet Epidemiol 2002;22(2):170–85. [PubMed: 11788962]

Zhao J, Jin L, Xiong M. Test for interaction between two unlinked loci. Am J Hum Genet 2006;79(5): 831–45. [PubMed: 17033960]

Zheng G, Chen Z. Comparison of maximum statistics for hypothesis testing when a nuisance parameter is present only under the alternative. Biometrics 2005;61(1):254–8. [PubMed: 15737101]

Table 1

Definitions for methods compared in the simulation studies

Method	Rank truncation points for gene-level summary	Rank truncation points for pathway level summary	
		Pathway with 12 genes	Pathway with 110 genes
ARTP-ARTP	$K_j = j\max\{1, [n/20]\}$, $1 \le j \le 5$	$K_j = j$, $1 \le j \le 12$	$K_j = 5j$, $1 \le j \le 10$
ARTP-RTP(q)	$K_j = j\max\{1, [n/20]\}$, $1 \le j \le 5$	$K_1 = q$ (1 truncation point)	$K_1 = q$ (1 truncation point)
RTP-ARTP	$K_1 = 1$ (1 truncation point)	$K_j = j$, $1 \le j \le 12$	$K_j = 5j$, $1 \le j \le 10$
ARTP-I	No gene level summary	$K_j = j$, $1 \le j \le 12$ (SNP-based)	$K_j = 5j$, $1 \le j \le 10$ (SNP-based)
ARTP-II	No gene-level summary	$K_j = 5j$, $1 \le j \le 12$ (SNP-based)	$K_j = 10j$, $1 \le j \le 10$ (SNP-based)
ARTP-III	No gene-level summary	$K_j = 10j$, $1 \le j \le 12$ (SNP-based)	$K_j = 20j$, $1 \le j \le 10$ (SNP-based)
RTP(q)	No gene-level summary	$K_1 = q$ (1 truncation point)	$K_1 = q$ (1 truncation point)

Table 2

Type I error under the significance level of 0.05. Results are estimated based on 2,000 replications under the 12 gene and 110 gene pathways

Methods*		12 Gene Pathway	110 Gene Pathway
Gene-based	RTP-ARTP	0.055	0.049
	ARTP-ARTP	0.054	0.048
SNP-based	ARTP-I	0.051	0.050
	ARTP-II	0.049	0.049
	ARTP-III	0.050	0.045

* The detailed definition for each method is given in Table 1.

Table 3

Power comparison between gene-based and SNP-based approaches. Results are estimated based on 500 replications under the pathway consisting of 12 genes (Pathway-12)

		Methods*							
		Gene-based		SNP-based					
Disease-associated gene size	Functional SNP typed	ARTP-ARTP	RTP-ARTP	ARTP-I	ARTP-II	ARTP-III	RTP(1)	RTP(5)	RTP(10)
Small	Yes	0.96	0.94	0.76	0.75	0.70	0.70	0.78	0.77
	No	0.90	0.86	0.58	0.52	0.46	0.55	0.57	0.53
Medium	Yes	0.94	0.93	0.87	0.90	0.90	0.85	0.88	0.89
	No	0.72	0.66	0.65	0.72	0.73	0.55	0.67	0.70
Large	Yes	0.87	0.84	0.89	0.93	0.92	0.83	0.89	0.91
	No	0.76	0.68	0.76	0.85	0.87	0.64	0.77	0.81

*The detailed definition for each method is given in Table 1 under the 12 Gene Pathway.

Table 4

Power comparison among various gene-based pathway analyses. Results are estimated based on 500 replications under the pathway consisting of 110 genes (Pathway-110)

	Gene-based methods*								
Number of disease-related genes (Odds ratio)	**ARTP-ARTP**	**ARTP-RTP(1)**	**ARTP-RTP(5)**	**ARTP-RTP(10)**	**ARTP-RTP(20)**	**ARTP-RTP(30)**	**ARTP-RTP(40)**	**ARTP-RTP(50)**	**ARTP-RTP(110)**
5 (1.3)	0.89	0.85	0.91	0.89	0.81	0.74	0.68	0.63	0.53
15 (1.2)	0.88	0.61	0.86	0.89	0.88	0.87	0.84	0.83	0.79
30 (1.15)	0.92	0.51	0.78	0.86	0.91	0.92	0.93	0.94	0.93
45 (1.13)	0.94	0.46	0.78	0.90	0.93	0.94	0.96	0.95	0.95

* The detailed definition for each method is given in Table 1 under the Pathway-110 column.

Table 5

Nicotinic receptors pathway definition

Gene name	Chromosome id	Number of SNPs included
CHRNB2	1	5
CHRNA1	2	8
CHRND	2	7
CHRNG	2	5
CHRNA9	4	13
CHRNA2	8	2
CHRNA6	8	5
CHRNB3	8	5
CHRNA10	11	10
CHRNA3	15	15
CHRNA5	15	8
CHRNA7	15	27
CHRNB4	15	12
CHRNB1	17	7
CHRNE	17	6
CHRNA4	20	7

Table 6

The significance level of association between the nicotinic receptors pathway and two smoking behaviors, cigarette per day (CPD) and duration of smoking (SMKDU)

Method*		CPD	SMKDU
Gene-based	RTP-ARTP	0.008	0.058
	ARTP-ARTP	0.013	0.049
SNP-based	ARTP-I	0.028	0.091
	ARTP-II	0.024	0.062
	ARTP-III	0.021	0.060

* The detailed definition for each method is given in Table 1 under the 12 Gene Pathway