PLOS | GENETICS

RESEARCH ARTICLE

A Powerful Procedure for Pathway-Based Meta-analysis Using Summary Statistics Identifies 43 Pathways Associated with Type II Diabetes in European Populations

Han Zhang[1], **William Wheeler**[2], **Paula L. Hyland**[1], **Yifan Yang**[3], **Jianxin Shi**[1], **Nilanjan Chatterjee**[4,5]*, **Kai Yu**[1]*

1 Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **2** Information Management Services Inc., Calverton, Maryland, United States of America, **3** Department of Statistics, University of Kentucky, Lexington, Kentucky, United States of America, **4** Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America, **5** Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America

* nchatte2@jhu.edu (NC); yuka@mail.nih.gov (KY)

OPEN ACCESS

Citation: Zhang H, Wheeler W, Hyland PL, Yang Y, Shi J, Chatterjee N, et al. (2016) A Powerful Procedure for Pathway-Based Meta-analysis Using Summary Statistics Identifies 43 Pathways Associated with Type II Diabetes in European Populations. PLoS Genet 12(6): e1006122. doi:10.1371/journal.pgen.1006122

Editor: Greg Gibson, Georgia Institute of Technology, UNITED STATES

Received: February 22, 2016

Accepted: May 20, 2016

Published: June 30, 2016

Data Availability Statement: Data from DIAbetes Genetics Replication And Meta-analysis (DIAGRAMv3) can be obtained from http://diagram-consortium.org/. Data are available from Genetic Epidemiology Research on Aging (GERA, dbGaP Study Accession: phs000674.v1.p1), http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000674.v1.p1. Data are available from Asian Genetic Epidemiology Network (AGEN) consortium, with contact email: yooncho33@hallym.ac.kr.

Abstract

Meta-analysis of multiple genome-wide association studies (GWAS) has become an effective approach for detecting single nucleotide polymorphism (SNP) associations with complex traits. However, it is difficult to integrate the readily accessible SNP-level summary statistics from a meta-analysis into more powerful multi-marker testing procedures, which generally require individual-level genetic data. We developed a general procedure called Summary based Adaptive Rank Truncated Product (sARTP) for conducting gene and pathway meta-analysis that uses only SNP-level summary statistics in combination with genotype correlation estimated from a panel of individual-level genetic data. We demonstrated the validity and power advantage of sARTP through empirical and simulated data. We conducted a comprehensive pathway-based meta-analysis with sARTP on type 2 diabetes (T2D) by integrating SNP-level summary statistics from two large studies consisting of 19,809 T2D cases and 111,181 controls with European ancestry. Among 4,713 candidate pathways from which genes in neighborhoods of 170 GWAS established T2D loci were excluded, we detected 43 T2D globally significant pathways (with Bonferroni corrected p-values < 0.05), which included the insulin signaling pathway and T2D pathway defined by KEGG, as well as the pathways defined according to specific gene expression patterns on pancreatic adenocarcinoma, hepatocellular carcinoma, and bladder carcinoma. Using summary data from 8 eastern Asian T2D GWAS with 6,952 cases and 11,865 controls, we showed 7 out of the 43 pathways identified in European populations remained to be significant in eastern Asians at the false discovery rate of 0.1. We created an R package and a web-based tool for sARTP with the capability to analyze pathways with thousands of genes and tens of thousands of SNPs.

Funding: The work of HZ, JS, and KY were supported by the Intramural Program of the National Institutes of Health and the National Cancer Institute. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: WW is employed by the Information Management Services Inc., which is a contract company working for NIH. The authors declare that no other competing interests exist.

Author Summary

As GWAS continue to grow in sample size, it is evident that these studies need to be utilized more effectively for detecting individual susceptibility variants, and more importantly, to provide insight into global genetic architecture of complex traits. Towards this goal, identifying association with respect to a collection of variants in biological pathways can be particularly insightful for understanding how networks of genes might be affecting pathophysiology of diseases. Here we present a new pathway analysis procedure that can be conducted using summary-level association statistics, which have become the main vehicle for performing meta-analysis of individual genetic variants across studies in large consortia. Through simulation studies we showed the proposed method was more powerful than the existing state-of-art method. We carried out a comprehensive pathway analysis of 4,713 candidate pathways on their association with T2D using two large studies with European ancestry and identified 43 T2D-associated pathways. Further examinations of those 43 pathways in 8 Asian studies showed that some pathways were trans-ethnically associated with T2D. This analysis clearly highlights novel T2D-associated pathways beyond what has been known from single-variant association analysis reported from largest GWAS to date. We also identify a novel locus for T2D in the European populations at chromosome 17q21 (rs1058018, $p = 3.06 \times 10^{-8}$).

Introduction

Genome-wide association study (GWAS) has become a very effective way to identify common genetic variants underlying various complex traits [1]. The most commonly used approach to analyze GWAS data is the single-locus test, which evaluates one single nucleotide polymorphism (SNP) at a time. Despite the enormous success of the single-locus analysis in GWAS, proportions of genetic heritability explained by already identified variants for most complex traits still remain small [2]. It is increasingly recognized that the multi-locus test, such as gene-based analysis and pathway (or gene-set) analysis, can be potentially more powerful than the single-locus analysis, and shed new light on the genetic architecture of complex traits [3, 4].

The pathway analysis jointly tests the association between an outcome and SNPs within a set of genes compiled in a pathway according to existing biological knowledge [4]. Although the marginal effect of a single SNP might be too weak to be detectable by the single-locus test, accumulated association evidence from all signal-bearing SNPs within a pathway could be strong enough to be picked up by the pathway analysis if this pathway is enriched with outcome-associated SNPs. Various pathway analysis procedures have been proposed in the literature, with the assumption that researchers could have full access to individual-level genotype data [5–9]. In practice, pathway analysis usually utilizes data from a single resource with limited sample size, as it can be challenging to obtain and manage individual-level GWAS data from multiple resources. As a result, pathway analysis often fails to identify new findings beyond what have already been discovered by the single-locus tests. To maximize the chance of discovering novel outcome-associated variants by increasing sample size, a number of consortia have been formed to conduct single-locus meta-analysis on data across multiple GWAS [10–14]. The single-locus meta-analysis aggregates easily accessible SNP-level summary statistics from multiple studies. Similarly, the pathway-based meta-analysis [15–21] that integrates the same type of summary data across participating studies could provide us a greater opportunity for detecting novel pathway associations. Future association studies focusing on identified pathways would have a much-reduced multiple-comparison burden in searching for novel variants with main or complicated nonlinear joint effects on the outcome of interest.

In this paper, we developed a pathway-based meta-analysis procedure by extending the adaptive rank truncated product (ARTP) pathway analysis procedure [9], which was originally developed for analyzing individual-level genotype data. The new procedure, called Summary based ARTP (sARTP), accepts input from SNP-level summary statistics, with their correlations estimated from a panel of reference samples with individual-level genotype data, such as the ones from the 1000 Genomes Project [22, 23]. This idea was initially used in conducting gene-based meta-analysis [24, 25] or conditional test [26]. As will be shown in the Results Section, sARTP usually has a power advantage over its competitors. In addition, sARTP is specifically designed for conducting pathway-based meta-analysis using SNP-level summary statistics from multiple studies. In real applications (e.g., the type 2 diabetes example described below), it is very common that different studies could have genotypes measured or imputed on different sets of SNPs. As a result, the sample size used in the pathway-based meta-analysis on each SNP can be quite different. Ignoring the difference in sample sizes across SNPs in a pathway-based meta-analysis would generate biased testing results.

Pathway analysis generally targets two types of null hypotheses [4], including the competitive null hypothesis [15, 16, 18–20], i.e., the genes in a pathway of interest are no more associated with the outcome than any other genes outside this pathway, and the self-contained null hypothesis [17, 21], i.e., none of the genes in a pathway of interest is associated with the outcome. The sARTP procedure focuses on the self-contained null hypothesis, as our main goal is to identify outcome-associated genes or loci. Also, as pointed out by [27], tests for the competitive null hypothesis often assume that genotype measured at different genes are independent when evaluating the association significance level. This assumption, which is generally invalid in practice, is unnecessary for sARTP when testing the self-contained null hypothesis. One may refer to [27] and [4] for more discussion and comparison of these two types of hypotheses.

The pathways defined in many public databases can consist of thousands of genes and tens of thousands of SNPs. To make the procedure applicable to large pathways, or pathways with high statistical significance, we implement sARTP with efficient and parallelizable algorithms, and adopt the direct simulation approach (DSA) [28] to evaluate the significance of the pathway association.

We demonstrated the validity and power advantage of sARTP through simulated and empirical data. We applied sARTP to conduct a pathway-based meta-analysis on the association between type 2 diabetes (T2D) and 4,713 candidate pathways defined in the Molecular Signatures Database (MSigDB) v5.0. The analysis used SNP-level summary statistics from two sources with European ancestry. One is generated from the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) consortium [13], which consists of 12,171 T2D cases and 56,862 controls across 12 GWAS. The other one is based on a T2D GWAS with 7,638 T2D cases and 54,319 controls that were extracted from the Genetic Epidemiology Research on Aging (GERA) study [29, 30]. The novel T2D-associated pathways detected in the European population were further examined in Asians using summary data generated by the Asian Genetic Epidemiology Network (AGEN) consortium meta-analysis, which combined 8 GWAS of T2D with a total of 6,952 and 11,865 controls from eastern Asian populations [10].

Materials and Methods

The Pathway-Based Meta-analysis Procedure

Here we describe the proposed method sARTP for assessing the association between a dichotomous outcome and a pre-defined pathway consisting of J genes. The same procedure can be applied to study a quantitative outcome with minor modifications.

Score statistics and their variance-covariance matrix. We assume we have data from L GWA studies, with each consisting of $n^{(l)}$ subjects, $l = 1,\cdots,L$. Each gene in that pathway can contain one or multiple SNP(s), while any two genes may have some overlapped SNPs. For simplicity, we use superscript l to represent an individual study. For subject i in study l, $i = 1,\cdots,n^{(l)}$, let $y_i^{(l)}$ be the dichotomous outcome (e.g., disease condition, case/control status) taking values from {0,1}, and let $X_i^{(l)}$ be the vector of covariates to be adjusted for. The centralized genotypes of q SNPs within a pathway are presented as a vector $G_i^{(l)} = (g_{i1}^{(l)}, \cdots, g_{iq}^{(l)})^T$ for subject i. We assume the following logistic regression model as the risk model

$$\operatorname{logit} \mathrm{P}(y_i^{(l)} = 1 | X_i^{(l)}, G_i^{(l)}) = (X_i^{(l)})^T \alpha^{(l)} + (G_i^{(l)})^T \gamma, \ \ i = 1, \cdots, n,$$

Under the self-contained null hypothesis H_0: $\gamma = 0$, we denote the maximum likelihood estimate of $\alpha^{(l)}$ as $\widehat{\alpha}^{(l)}$. Let $\widehat{y}_i^{(l)} = 1/(1 + \exp(-X_i^{(l)}\widehat{\alpha}^{(l)}))$ and $u_i^{(l)} \ = \widehat{y}_i^{(l)}(1 - \widehat{y}_i^{(l)})$. The Rao's score statistic vector on γ, which is the sum of score vectors from L participating studies, follows the asymptotic multivariate normal distribution $N(0,V)$, where

$$S = (S_t)_{q\times 1} = \sum_{l=1}^{L} \sum_{i=1}^{n^{(l)}} G_i^{(l)}(y_i^{(l)} - \widehat{y}_i^{(l)}) \tag{1}$$

and

$$\begin{aligned} V = \sum_{l=1}^{L} \Bigg(& \sum_{i=1}^{n^{(l)}} u_i^{(l)} G_i^{(l)} (G_i^{(l)})^T \\ & - \sum_{i=1}^{n^{(l)}} u_i^{(l)} G_i^{(l)} (X_i^{(l)})^T \left(\sum_{i=1}^{n^{(l)}} u_i^{(l)} X_i^{(l)} (X_i^{(l)})^T \right)^{-1} \sum_{i=1}^{n^{(l)}} u_i^{(l)} X_i^{(l)} (G_i^{(l)})^T \Bigg). \end{aligned} \tag{2}$$

For study l, let $n_t^{(l)}$ be the number of subjects having their genotypes measured as $H_t^{(l)}$ (or imputed) at SNP t, where $H_t^{(l)} = (g_{1t}^{(l)}, \cdots, g_{n_t^{(l)}t}^{(l)})^T$. As pointed out by Hu, Berndt (24) if the covariates and genotypes are uncorrelated or weakly correlated, the covariance between scores at SNPs t and S can be approximated as

$$\begin{aligned} V_{ts} & \approx \sum_{l=1}^{L} n_{ts}^{(l)} \bar{u}^{(l)} \widehat{\mathrm{Cov}(H_t^{(l)}, H_s^{(l)})} \\ & \approx \sum_{l=1}^{L} n_{ts}^{(l)} \rho_{ts} \sqrt{\bar{u}^{(l)} \widehat{\mathrm{Var} H_t^{(l)}}} \sqrt{\bar{u}^{(l)} \widehat{\mathrm{Var} H_s^{(l)}}}, \ \ t, s = 1, \cdots, q, \end{aligned} \tag{3}$$

where $n_{ts}^{(l)}$ is the number of samples that have their genotypes available at both SNPs in study l, $\bar{u}^{(l)} = (n^{(l)})^{-1} \sum_{i=1}^{n^{(l)}} u_i^{(l)}$, and $\widehat{\mathrm{Cov}(H_t^{(l)}, H_s^{(l)})} = (n^{(l)})^{-1} \sum_{i=1}^{n^{(l)}} g_{it}^{(l)} g_{is}^{(l)}$. Here, we assume that the Pearson's correlation coefficient ρ_{ts} between two SNPs is the same among all participating studies. This assumption is valid as long as subjects from all studies are sampled from the same source population, or the population under study is relatively homogeneous, such as a study of subjects with European ancestry in the United States.

When only the summary statistics, i.e., the estimated marginal log odds ratios $\widehat{\beta}_t^{(l)}$ and their standard errors $\tau_t^{(l)}$ are available for each of the L studies, the score statistic at SNP t, defined by (Eq 1) can be approximated as

$$S_t \approx \sum_{l=1}^{L} \left(\tau_t^{(l)}\right)^{-2} \widehat{\beta}_t^{(l)};\ t = 1, \cdots, q. \tag{4}$$

Note that $n_t^{(l)} \bar{u}^{(l)} \widehat{\mathrm{Var} H_t^{(l)}} \approx \left(\tau_t^{(l)}\right)^{-2}$, thus according to (Eq 3), we have

$$V_{ts} \approx \sum_{l=1}^{L} \frac{n_{ts}^{(l)}}{\sqrt{n_t^{(l)} n_s^{(l)}}} \frac{\rho_{ts}}{\tau_t^{(l)} \tau_s^{(l)}}. \tag{5}$$

Assume that ρ_{ts} can be estimated from a public dataset (e.g., 1000 Genomes Project) and the sample sizes $n_t^{(l)}$ and $n_{ts}^{(l)}$ are known, we can approximately recover the variance-covariance matrix $V = (V_{ts})_{q\times q}$ of score statistics $S = (S_t)_{q\times 1}$. In cases when we only have the SNP p-value p and its marginal log odds ratio $\widehat{\beta}$, we can compute its standard error as $\tau = |\widehat{\beta}|/\sqrt{\chi^2_{1,p}}$, where $\chi^2_{1,p}$ is the quantile satisfying $\mathrm{P}(\chi^2_1 \geq \chi^2_{1,p}) = p$, with χ^2_1 representing a 1-df chi-squared random variable.

Combining score statistics for pathway analysis. With recovered score statistics vector S and its variance-covariance matrix V, we can conduct a pathway association test using the framework of the ARTP method. The ARTP method first combines p-values of individual SNPs within a gene to form a gene-based association statistic (i.e., the gene-level p-value), and then combines the gene-level p-values into a final testing statistic for the pathway-outcome association. In the original ARTP method, [9] proposed the use of a resampling-based method to evaluate the significance level of the pathway association test. Here we integrate the SNP-level score statistics into the ARTP framework and use DSA [28] to evaluate the significance level, which is much faster than the original ARTP algorithm [31]. Below is a brief summary of the improved ARTP algorithm.

First we obtain the p-values $p_{t_1}^{(0)}, \cdots, p_{t_{q_j}}^{(0)}$ of q_j distinct SNPs in gene j as $p_t^{(0)} = \mathrm{P}(\chi^2_1 \geq S_t^2/V_{tt})$. Let $p_{j(1)}^{(0)}, \cdots, p_{j(q_j)}^{(0)}$ be their order statistics such that $p_{j(1)}^{(0)} \leq \cdots \leq p_{j(q_j)}^{(0)}$. For any predefined integer K and SNP-level cut points $c_1 < \cdots < c_k$, we define the observed negative log product statistics for that gene at cut point c_k as

$$w_{jk}^{(0)} = -\sum_{t=1}^{\min(q_j, c_k)} \log p_{j(t)}^{(0)},\ \ k = 1, \cdots, K.$$

We sample M copies of vectors of the score statistic from the null distribution $N(0,V)$ and convert each of them to be the tail probability of χ^2_1 as $p_{t_1}^{(m)}, \cdots, p_{t_{q_j}}^{(m)}$, $m = 1,\cdots,M$, which are then used to calculate $w_{jk}^{(m)}$, $m = 1,\cdots,M$. The significance of $w_{jk}^{(0)}$ can be estimated as

$$\xi_{jk}^{(0)} = \frac{\#\left\{w_{jk}^{(m)} \geq w_{jk}^{(0)}; m = 1, \cdots, M\right\}}{M+1}.$$

The ARTP statistic for testing association between gene j and the outcome is defined as $T_j^{(0)} = \min_{k=1,\cdots,K} \xi_{jk}^{(0)}$. Note that for any $w_{jk}^{(m)}$, the set $\{w_{jk}^{(m\prime)} : m' \in \{0, \cdots, M\}$ and $m' \neq m\}$ forms its empirical null distribution. The significance of $w_{jk}^{(m)}$ therefore can be estimated as

$$\xi_{jk}^{(m)} = \frac{\#\left\{w_{jk}^{(m')} \geq w_{jk}^{(m)};\ m' \neq m \text{ and } m' = 1,\cdots,M\right\}}{M+1},\ m = 1,\cdots,M.$$

This idea, which was given by [32], can be used to avoid the computationally challenging nested two-layer resampling procedure for evaluating p-values. The p-value of $T_j^{(0)}$ can be readily calculated as

$$z_j^{(0)} = \frac{\#\left\{T_j^{(m)} \leq T_j^{(0)} : m = 1,\cdots,M\right\}}{M+1},\ j = 1,\cdots,J.$$

where $T_j^{(m)} = \min_{k=1,\cdots,K} \xi_{jk}^{(m)}$. $z_j^{(0)}$ is the estimated gene-level p-value for the association between the outcome and the *j*th gene. To obtain the pathway p-value, a similar procedure as above can be applied to combine already established gene-level p-values $z_j^{(0)}$, $j = 1,\cdots,J$, through a set of K' gene-level cut points $d_1 < \cdots < d_{K'}$. For simplicity, let $\zeta_k^{(0)}$ be the significance (p-value) of negative log product statistics defined on $z_j^{(0)}$, $j = 1,\cdots,J$ at a specific cut point d_k, $k = 1,\cdots,K'$. The ARTP statistic for the pathway association is defined as $T^{(0)} = \min_{k=1,\cdots,K'} \zeta_k^{(0)}$. The top d_{k^*} genes, at which $\zeta_{k^*}^{(0)} = \min_{k=1,\cdots,K'} \zeta_k^{(0)}$, can be regarded as the set of selected candidate genes that collectively convey the strongest pathway association signal.

In the following discussion, we will use the term sARTP to represent the proposed pathway analysis procedure using the SNP-level summary statistics as input, and reserve the term ARTP to represent the original ARTP procedure that requires the individual-level genetic data. Both procedures adopt the DSA algorithm to accelerate evaluating the significance level. When performing the pathway analysis in this paper, we set SNP-level cut points as $(c_1,c_2) = (1,2)$, i.e., gene-level association is summarized by one or two most significant SNPs within each gene, and gene-level cut points as $d_k = k\max(1, \lceil J/20 \rceil)$, $k = 1,\cdots,10$, where J is the number of genes in a pathway, and $\lceil J/20 \rceil$ is the largest integer that is less or equal to J / 20. We used $M = 10^5$ DSA steps to assess the significance level of each pathway in the initial screening. For pathways with estimated p-values $<10^{-4}$, we further refined their p-value estimates with $M = 10^7$ or 10^8 DSA steps.

Applying sARTP to meta-analysis result. Many GWAS consortia usually publish their meta-analysis results by providing only the combined results from the fixed effects model, rather than the summary statistics from each participating study. We can apply sARTP to this meta-analysis result directly, with some modifications. First, since the reported marginal log odds ratios for each SNP by using the fixed effects inverse-variance weighting method is given by

$$\widehat{\beta}_t = \frac{\sum_{l=1}^{L} \left(\tau_t^{(l)}\right)^{-2} \widehat{\beta}_t^{(l)}}{\sum_{l=1}^{L} \left(\tau_t^{(l)}\right)^{-2}},$$

with its standard error given by

$$\tau_t = \left(\sum_{l=1}^{L} (\tau_t^{(l)})^{-2} \right)^{-1/2}. \tag{6}$$

Based on (Eq 4), we can see $S_t \approx \tau_t^{-2}\widehat{\beta}_t$. By assuming large sample sizes and certain conditions (see S1 Text), we can also approximate the covariance between S_t and S_s, which is given by (Eq 5), as

$$V_{ts} \approx \frac{n_{ts}}{\sqrt{n_t n_s}} \frac{\rho_{ts}}{\tau_t \tau_s}, \tag{7}$$

where $n_t = \sum_{l=1}^{L} n_t^{(l)}$, and $n_{ts} = \sum_{l=1}^{L} n_{ts}^{(l)}$. Thus, using just the meta-analysis result, without knowing summary statistics from each participating study, we can still obtain S_t exactly, and approximately recover V_{ts}. As a result, we can carry out the pathway-based meta-analysis based on the SNP-level meta-analysis result as if it were summary data from a single study. We call this approach the Meta-analysis based sARTP (MsARTP).

However, to apply the MsARTP, we need additional sample size information n_t and n_{st} in order to properly estimate the variance-covariance matrix defined by (Eq 7). If the same set of SNPs are studied by all participating studies, we have $n_t = n_s = n_{st}$, and the approximation (Eq 7) becomes $V_{ts} \approx \frac{\rho_{ts}}{\tau_t \tau_s}$, i.e., we can obtain the estimated variance-covariance matrix without knowing n_t and n_{st}. But in most applications, not all GWAS choose the same SNP genotyping array, even after the imputation using the same reference genomes. As a result, the SNP coverage, i.e., the set of SNPs evaluated in each participating study can be quite different. In those situations, we need to know the SNP coverage information in each participating study in order to obtain n_s and n_{st}. We will show in the Results Section that using MsARTP with an inappropriate uniform coverage assumption (i.e., $n_{ts} = n_t = n_s$), which is commonly made by many multilocus approaches, can lead to inflated type I error.

Given SNP-level summary statistics from each participating study, we can either apply sARTP directly, or first conduct a SNP-level meta-analysis, and then apply MsARTP to the meta-analysis result. These two approaches use the same score statistics, and different but consistent estimates for the variance-covariance matrix. Numeric experiments in the Results Section suggest that these two approaches generate vary similar pathway p-values.

Study Materials

Pathway and gene definition. We downloaded definitions for 4,716 human and murine (mammalian) pathways (gene sets) from the MSigDB v5.0 (C2: curated gene sets). Genomic definitions for genes were downloaded from Homo sapiens genes NCBI36 and reference genome GRCh37.p13 using the Ensemble BioMart tool.

DIAGRAM study. The DIAGRAM (DIAbetes Genetics Replication And Meta-analysis) consortium conducted a large-scale GWAS meta-analysis to characterize the genetic architecture of T2D [13]. We downloaded the summary statistics generated by the DIAGRAMv3 (Stage 1) GWAS meta-analysis from www.diagram-consortium.org [13]. The meta-analysis studied 12 GWAS with European ancestry consisting of 12,171 cases and 56,862 controls. Up to 2.5 million autosomal SNPs with minor allele frequencies (MAFs) larger than 1% were imputed using CEU samples from Phase II of the International HapMap Project. Study-specific covariates were adjusted in testing T2D-SNP association under an additive logistic regression model [13]. SNP-level summary statistics from each GWAS were first adjusted for residual population structure using the genomic control (GC) method [33], and then combined in the fixed effects meta-analysis.

We sorted 2.5 million autosomal SNPs by their corresponding meta-analysis sample sizes in S1 Fig, which shows that there are two major groups of SNPs with equal sample sizes. One group of 469,985 SNPs (19.0%) had 12,171 cases and 56,862 controls, which included all the available samples in the meta-analysis; another group of 1,431,361 SNPs (57.9%) had 9,580 cases and 53,810 controls. Since the calculation of covariance V_{ts} in (Eq 7) relies on n_{ts}, the number of samples having genotypes available at both SNP s and SNP t, in order to obtain an accurate estimate of n_{ts}, we focused on these two groups of SNPs, which in combination had a total of 1,901,346 SNPs. For any two SNPs in this reduced set, it is certain $n_{ts} = \min(n_t, n_s)$. The Pearson's correlation coefficients ρ_{ts} were estimated using an external reference panel consisting of genotypes on 503 European subjects (CEU, TSI, FIN, GBR, and IBS) from the 1000 Genomes Project (Phase 3, v5, 2013/05/02).

GERA study. We assembled a GWAS on T2D from the Genetic Epidemiology Research on Adult Health and Aging (GERA, dbGaP Study Accession: phs000674.v1.p1). The GERA project includes a cohort of over 100,000 adults who are members of the Kaiser Permanente Medical Care Plan, Northern California Region, and participating in the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH). From the GERA data, we compiled a GWAS with 7,638 T2D cases and 54,319 controls (subjects without T2D) who self-reported to be non-Hispanic White Europeans in the RPGEH survey. We performed the genotype imputation with IMPUTE2 [34] using CEU reference samples from Phase II of the International HapMap Project. After removing SNPs with low imputation quality ($r^2 < 0.3$), we ended up with 2.4 million SNPs for further analysis. In the single-locus analysis, we adjusted for the categorized body mass index (BMI) provided in the downloaded dataset (adding a category for missing BMI), gender, year of birth (in five-year categories), a binary indicator on whether or not a participant was diagnosed with cancer (includes malignant tumors, neoplasms, lymphoma and sarcoma), and the top five eigenvectors for the adjustment of population stratification. In the following discussion, we refer this assembled T2D GWAS as the GERA study.

When analyzing the SNP-level summary data from the GERA study, the Pearson's correlation coefficients ρ_{ts} were estimated using an external reference panel consisting of genotypes on 503 European subjects from the 1000 Genomes Project.

AGEN-T2D study. The Asian Genetic Epidemiology Network (AGEN) consortium carried out a meta-analysis by combining eight GWAS of T2D with a total of 6,952 cases and 11,865 controls from eastern Asian populations [10]. The meta-analysis was conducted with the fixed effect model. We obtained SNP-level summary statistics on 2.6 million imputed and genotyped autosomal SNPs from AGEN, and used this summary data to evaluate whether pathway associations identified in European populations remain to be present in Asians. We adopted an external reference panel consisting of 312 eastern Asian subjects (103 from CHB, 105 from CHS, and 104 from JPT) from the 1000 Genomes Project for the variance-covariance matrix estimation in the pathway analysis.

Results

Simulation Studies

Firstly, we conducted a simulation study to evaluate the empirical size of sARTP and MsARTP. Secondly, we compared empirical powers of different strategies for carrying out pathway-based meta-analysis that integrated summary statistics from multiple studies. We also evaluated whether results from sARTP were consistent with the ones from MsARTP. Thirdly, we compared our method to the recently developed method aSPUsPath [8] that can be used for pathway-based meta-analysis. We used the R package, aSPU (version 1.39), with the default settings given in [8, 17] to conduct the aSPUsPath test.

Empirical size of sARTP and MsARTP. To evaluate the empirical size of sARTP and MsARTP, we conducted a simulation study by using individual-level GWAS data of the pathway PUJANA_BREAST_CANCER_WITH_BRCA1_MUTATED_UP (including 728 SNPs in 50 genes) from the GERA study. We picked 12,000 samples randomly for this experiment. By keeping their genotypes unchanged, we randomly assigned 6,000 subjects as cases and the remaining as controls to generate 500,000 datasets. We split each dataset into three case-control studies, each with 2,000 cases and 2,000 controls. To mimic the scenario when not all studies have their genotypes measured on the same set of SNPs (such as the one occurred in the DIAGRAM and AGEN data), we assumed that each case-control study had genotypes measured on only half of SNPs in the pathway. For each generated dataset that consisted of three case-control studies, we applied sARTP to the SNP-level summary data obtained from each case-control study, and MsARTP to the meta-analysis result based on the three case-control studies, with the variance-covariance matrix estimated by an external reference panel (with 503 European reference samples from the 1000 Genomes Project), or an internal reference panel (with 500 samples randomly selected from the GERA data).

Based on results from the 500,000 generated datasets, this simulation study showed that both sARTP and MsARTP, using the internal or external reference samples, can well control their empirical sizes (Table 1). Given the same reference panel, the p-values estimated from sARTP and MsARTP are highly consistent (Pearson's correlation coefficient $>$ 0.99). Furthermore, the p-values of sARTP (or MsARTP) estimated with an external or internal reference panel are also very consistent (Pearson's correlation coefficient $>$ 0.99). More numeric experiments demonstrating the validity of sARTP under the null are described in S2 Text.

To demonstrate the importance of knowing n_t and n_{ts} when applying MsARTP to the meta-analysis result, we analyzed each simulated dataset using MsARTP assuming the uniform coverage ($n_{ts} = n_t = n_s$). We called this approach MsARTP-u. It is clear from Table 1 that MsARTP-u assuming the uniform coverage suffers from inflated type I errors with either the internal or external reference panel.

Empirical power of sARTP and MsARTP for pathway-based meta-analysis. We conducted a set of simulation studies to compare the power of different strategies to carry out pathway analysis when SNP-level summary statistics were available from multiple studies. We considered a hypothetical pathway consisting of 50 genes randomly selected from chromosome

Table 1. Empirical sizes of the sARTP, MsARTP, and MsARTP-u procedures.

	Reference	Size				
		0.05	0.01	0.005	0.001	0.0005
sARTP	External[a]	0.050	0.0093	0.0040	0.00078	0.00044
	Internal[b]	0.046	0.0087	0.0042	0.00074	0.00040
MsARTP	External	0.048	0.0093	0.0040	0.00076	0.00041
	Internal	0.048	0.0084	0.0042	0.00074	0.00041
MsARTP-u	External	0.082	0.018	0.0081	0.0013	0.00064
	Internal	0.094	0.022	0.011	0.0016	0.00081

Empirical sizes are estimated based on 500,000 datasets simulated from the GERA data.

[a] Using 503 European samples from the 1000 Genomes Project as an external reference;

[b] Using 500 samples from the GERA data as an internal reference.

doi:10.1371/journal.pgen.1006122.t001

17, each with 20 randomly chosen SNPs. The joint genotype distribution at the 20 SNPs within each gene was defined by the observed genotypes in the GERA study. We further assumed that all genes in that pathway are independent. This assumption is unnecessary for sARTP and MsARTP, but it was introduced for simplifying the simulation. For the risk model, we assumed the first $\mathcal{M}$ ($\mathcal{M} = 5, 10, 15$) genes were associated with the outcome. Within each outcome-associated gene, we picked the SNP with its MAF closest to the median MAF level within the gene to be functional. We considered the following risk model

$$\text{logit}\, \mathrm{P}(y = 1 \mid g_1^*, \cdots, g_{\mathcal{M}}^*) = \alpha + \sum_{l=1}^{\mathcal{M}} \gamma_l^* g_l^*, \tag{8}$$

where g_l^* is the genotype (encoded as 0, 1, or 2 according to counts of minor alleles) at the functional SNP within gene l. Under this model, γ_l^* is also the marginal log odds ratio for the lth functional SNP [9]. Given the sample sizes of cases and controls, and the MAF of the lth functional SNP, γ_l^* was chosen such that the theoretical power of the trend test to detect the lth functional SNP is equal to $\mathcal{P}$ ($\mathcal{P} = 0.3, 0.4$), with 0.05 as the targeted type I error rate. For every pair of $(\mathcal{M},\ \mathcal{P})$, we generated 1,000 datasets, each consisting of three case-control studies, with the same sample size and SNP coverage configurations used for evaluating the empirical size. Given the genotype distribution in the general population, individual-level genotype data for a case-control study can be generated according to the assumed risk (model 8).

We assumed that only SNP-level summary statistics from each of the three studies were available. For each simulated dataset, we applied sARTP and MsARTP, using either an internal or external reference panel to estimate the variance-covariance matrix. The sARTP and MsARTP approaches integrate association evidence across SNP-level summary statistics, which are obtained by pooling information from all participating studies on individual SNPs. As a comparison, we also considered a naïve approach, in which we first applied sARTP to analyze the summary statistics from each study separately, and then combined the three pathway p-values with Fisher's method. This naïve approach could be useful when the researchers do not have access to the SNP-level summary data but the pathway p-values from individual studies. The empirical powers are compared at the type I error level of 0.05, and are summarized in Table 2. It is obvious that the pathway-based meta-analysis using sARTP, with either the internal or external reference panel, have almost the same level of power as the MsARTP method. It is also evident that both sARTP and MsARTP are more powerful than the naïve approach, which suggests that it is always be beneficial to have the SNP-level summary statistics from each participating study, or SNP-level meta-analysis result when conducting a pathway analysis.

Given the SNP coverage information, the MsARTP method is a valid pathway association test that has well controlled type I error and similar power to the sARTP method. In the following analysis, either sARTP or MsARTP is chosen depending on the type of available data. For the sake of simplicity, we always label the chosen procedure as sARTP.

Power comparison between sARTP and aSPUsPath. Since the aSPUsPath method in the current aSPU package cannot handle summary data from multiple studies, or meta-analysis results from studies with varied SNP coverage, we focused on the scenario with just one study, and adopted the similar simulation strategy as the one used by [8] to compare the power between sARTP and aSPUsPath. We simulated haplotypes on a set of SNPs within a gene in the general population using the algorithm of Wang and Elston [35]. Then the joint genotypes on a subject can be formed by randomly pairing two haplotypes. In brief, we first chose the MAF for each SNP by randomly sampling a value from the uniform distribution $U(0.1,0.4)$. Then for the set of SNPs in a gene we sampled a latent vector $Z = (z_1, \cdots, z_q)^T$ from a

Table 2. Power comparisons under the type I error rate of 0.05 when analyzing data from three studies.

$\mathcal{P}$[a]	$\mathcal{M}$[b]	Internal reference[c]			External reference[d]		
		sARTP	MsARTP	Fisher	sARTP	MsARTP	Fisher
0.3	5	0.165	0.170	0.110	0.170	0.167	0.105
	10	0.405	0.402	0.229	0.399	0.401	0.221
	15	0.573	0.578	0.334	0.564	0.561	0.323
0.4	5	0.292	0.293	0.162	0.295	0.297	0.154
	10	0.642	0.637	0.363	0.640	0.635	0.362
	15	0.858	0.858	0.574	0.855	0.856	0.561

For every pair of $\mathcal{P}$ and $\mathcal{M}$, the empirical powers are computed from 1,000 simulated datasets at the level of 0.05. Each dataset contains three studies. The pathway consists of 50 independent genes, each with 20 SNPs. Fisher's method is used to combine the three pathway p-values obtained by applying sARTP to the SNP-level summary data from each of three studies separately.

[a]The theoretical power of the single-locus trend test on the functional SNP under the type I error rate of 0.05, given the sample sizes of cases and controls, and the MAF of the functional SNP;

[b]The number of genes including the functional SNPs;

[c] Using 500 samples from the GERA data as an internal reference;

[d]Using 503 European samples from the 1000 Genomes Project as an external reference.

doi:10.1371/journal.pgen.1006122.t002

multivariate normal distribution with a covariance matrix Cov(z_i, z_j) = $\rho^{|i-j|}$,$1\leq i,j\leq q$, where ρ was sampled from the uniform distribution $U(0,0.8)$ for a given gene. We randomly picked 50% of the SNPs and converted their simulated z_i into minor and major alleles (coded as 0, 1), with the cuts chosen for each z_i such that the resultant minor allele has its frequency defined by the specified MAF. For the remaining SNPs, we used the same algorithm to dichotomize $-z_i$ into minor and major alleles. This created a more realistic haplotype structure such that a haplotype can consist of a mixture of minor and major alleles. Genotypes on SNPs from different genes were generated independently.

Given the number of genes (20, 50, or 80) in a pathway, the proportion of genes (5%, 10%, 20%, and 30%) associated with the outcome, and a chosen common value for all log odds ratios (γ^*) in the risk (model 8), we repeated the following steps to generate 1,000 case-control studies, with each consisting of 1,000 cases and 1,000 controls. First, the number of SNPs within each gene was randomly chosen from 10 to 100. Second, for each randomly selected outcome-associated gene, we randomly picked a functional SNP. Third, we use the aforementioned algorithm of Wang and Elston [35] to generate the individual-level genotype data for a case-control study according to the specified risk model. We also considered the situation where all γ^* in the risk (model 8) had the same magnitude but different directions. More precisely, when generating a case-control study at the third step, we defined the risk (model 8) by randomly choosing the direction of each log odds ratio to be positive or negative with equal probability. Furthermore, we considered a more complex scenario where each outcome-associated gene had one or two functional SNPs, each with equal probability.

All simulation results are given in S1 and S2 Tables. It is clear that sARTP are generally more powerful than aSPUsPath, especially when the signal-to-noise ratio (the proportion of genes including a functional SNP) is relatively low. The two types of tests tend to have comparable performance when the signal-to-noise ratio increases to 30%, although it is uncommon for a candidate pathway to have such a high signal-to-noise ratio in real applications. For example, among the 4,713 candidate pathways analyzed in the next section, only 4.2% and 0.9% of the pathways have over 20% and 30% of their genes that are likely to contain association signals (i.e., with gene-level p-values < 0.05).

From S1 and S2 Tables, we also notice that the advantage of sARTP over aSPUsPath is more evident if not all minor alleles of the functional SNPs are deleterious (or protective) variants (i.e., γ^* in the risk (model 8) are not all positive). This is expected, as the sARTP approach does not take the effect direction of the minor allele at each SNP into consideration, while aSPUsPath integrates a set of candidate statistics, including the one similar to the burden test that assumes all minor alleles are either deleterious or protective. When this assumption is not valid, the inclusion of the burden test statistic in aSPUsPath is unlikely to enhance the power, but certainly would increase the multiple-testing penalty.

Evaluation of sARTP Using Data from T2D Studies

To demonstrate the consistency between results obtained by sARTP using SNP-level summary statistics and the ones by ARTP using individual-level genotype data, we compared pathway analysis results from three different procedures on the 4,713 candidate pathways using the GERA GWAS data. Details on how those 4,713 pathways were pre-processed are given in the Results of T2D Pathway Analysis Section. We applied sARTP to the SNP-level summary statistics generated from the GERA study, using either an internal or an external reference panel. We also obtained the pathway p-values by directly applying the ARTP method to the individual-level GERA GWAS data. Fig 1 shows the comparison among p-values from these three analyses, and demonstrates that all three approaches can generate very consistent results.

Results of T2D Pathway Analysis

Findings from the European populations. Since our goal was to identify new susceptibility loci for T2D through the pathway analysis, we excluded 170 high evidence T2D associated SNPs that were either listed in [13] or found from the GWAS Catalog satisfying the following

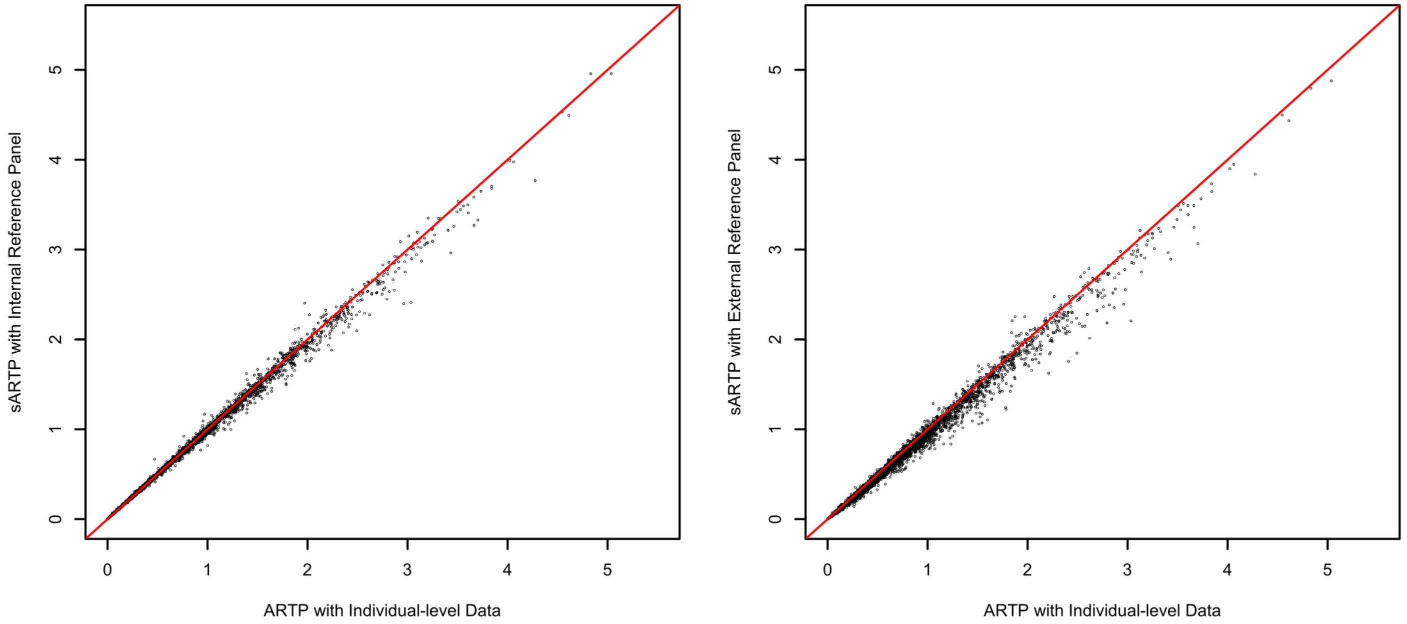

Fig 1. Comparisons of p-values from three types of pathway analyses on the GERA data. Based on the GERA data, 4,713 pathways are analyzed in three different ways. Pathway p-values obtained by ARTP using the GERA individual-level genetic data (x-axis) are compared with the ones obtained by sARTP using summary statistics in combination with the internal reference panel that consists of 500 randomly selected GERA samples (left), and the ones using the summary statistics in combination with the external reference panel that consists of 503 European subjects from the 1000 Genomes Project (right).

doi:10.1371/journal.pgen.1006122.g001

three conditions simultaneously: (1) were investigated by GWAS of samples with European ancestry; (2) had reported p-values $<10^{-7}$ on the initial study; and (3) were replicated on independent studies. We excluded 195 SNPs that has their single-locus testing p-values less than 10^{-7} in either DIAGRAM or GERA data to ensure that the pathway analysis result was not driven by a single SNP. In addition, we further excluded genes within a ±500kb region from each of the removed SNPs to eliminate potential association signals that could be caused by linkage disequilibrium (LD) with the index SNPs.

We conducted three types of pathway-based meta-analyses using sARTP, including the one using the DIAGRAM SNP-level summary statistics, the one using the GERA SNP-level summary statistics, and the pathway meta-analysis combining SNP-level summary statistics from both DIAGRAM and GERA studies. When applying the pathway-based meta-analysis to a single gene, we refer to this as the gene-level meta-analysis. We used the external reference panel of 503 Europeans from the 1000 Genomes Project to estimate the variance-covariance matrix.

Before performing a pathway analysis, we applied LD filtering to remove redundant SNPs. For any two SNPs with their pairwise squared Pearson's correlation coefficient > 0.9 estimated from the external reference panel from the 1000 Genomes Project, we removed the one with a smaller value defined as, $2f(1{-}f)n_0n_1n^{-1}$, where n_0 and n_1 are numbers of controls and cases, $n = n_0 + n_1$, and f is the MAF based on the reference panel. This value is proportional to the non-centrality parameter of the trend test statistic at a given SNP. We also excluded SNPs with MAF $<$ 1%. After all SNP filtering steps, we had a total of 4,713 pathways for the analysis. The summary of the number of genes and SNPs used in each pathway analysis is given in S2 Fig.

The DIAGRAM study had a genomic control inflation factor $\lambda_{GC} = 1.10$ based on the published meta-analysis result. The assembled GERA T2D GWAS had $\lambda_{GC} = 1.08$. When conducting the pathway analysis on each of two studies, we adjusted the inflation by using the corresponding $\sqrt{\lambda_{GC}}$ to rescale the standard error of estimated log odds ratio at each SNP. The single-locus meta-analysis combining results from DIAGRAM and GERA datasets had an inflation factor $\lambda_{GC} = 1.067$ after each study had adjusted for its own inflation factor. We further adjusted this inflation in the pathway and gene-level meta-analysis when combining SNP-level summary statistics from both studies using formulas (4) and (5).

The Q-Q plots of gene-level and pathway p-values are given in Fig 2. Gene-level p-value Q-Q plots based on the three analyses show no sign of inflation with their λ_{GC} close to 1.0, but suggest that there are enriched gene-level association signals at the tail end. The pathway p-value Q-Q plots, on the other hand, shift away from the diagonal identify line and have much higher λ_{GC}, which suggests that T2D associated genes are preferably included in pathways under study. In fact, it can be seen from S3 Fig that a gene with a smaller gene-level meta-analysis p-value tends to be included in more pathways, even though the 4,713 pathways collected from MSigDB v5.0 are not specifically defined for the study of T2D.

Fig 2 illustrates that the gene and pathway level signal from the GERA study tends to be slightly stronger than that from the DIAGRAM study. The main reason is that the DIAGRAM summary result had gone through two rounds of inflation adjustments, with the first round done at each participating study, and the second round on the meta-analysis result. Also, its second round adjustment ($\lambda_{GC} = 1.10$) is larger than the one applied to the GERA study ($\lambda_{GC} = 1.08$). Adjusting for λ_{GC} in the pathway analysis could be too conservative, since some proportion of the inflation can be caused by the real polygenic effect. A less conservative adjustment could be possible, but it might not be adequate. More discussions on this issue are given in the Discussion Section.

Based on the pathway meta-analysis on a total of 4,713 pathways, we identified 43 significant pathways with p-values less than $1.06{\times}10^{-5}$, the family-wise significant threshold based

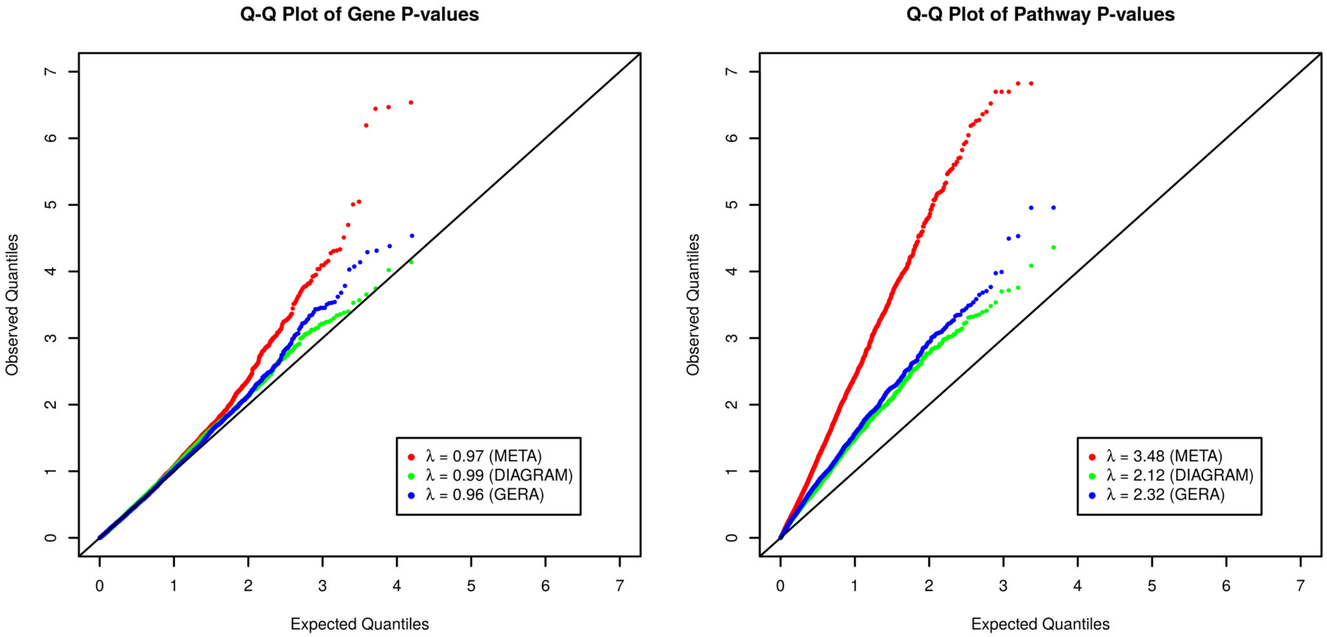

Fig 2. Q-Q plots of gene-level and pathway p-values based on the sARTP procedure on the DIAGRAM study, the GERA study, and the two studies combined. (Left) Q-Q plots of gene-level p-values on 15,946 genes based on the sARTP gene-based analysis of the DIAGRAM study (DIAGRAM), the GERA study (GERA), and the two studies combined (META). (Right) Q-Q plots of pathway p-values on 4,713 pathways based on the sARTP pathway analysis of the DIAGRAM study (DIAGRAM), the GERA study (GERA), and the two studies combined (META).

doi:10.1371/journal.pgen.1006122.g002

on the Bonferroni correction. Their pathway meta-analysis results as well as results from individual studies are summarized in Table 3. More detailed results on each of 43 significant pathways are given in the S6–S48 Figs and S6 Table. There are a total of 15,946 unique genes in all 4,713 pathways. The top 50 genes with smallest gene-level p-values based on the gene meta-analysis are listed in S3 Table. Because of the LD filtering, a gene belonging to two pathways might end up with slightly different sets of SNPs. To remove this ambiguity, we obtained the gene-level p-values by conducting a gene-level meta-analysis on each gene separately.

From Table 3, we can notice that some identified pathways have relatively weak association signals from each of the two studies, but have very significant p-values based on the pathway meta-analysis on the two studies combined. For example, the pathway RIZ_ERYTHROID_DIFFERENTIATION has p-values of 0.0233 and 0.0231 based on DIAGRAM and GERA studies, respectively. Combining these two p-values using Fisher's method yields a p-value of 0.0046. On the other hand, the pathway meta-analysis produces a much more significant result ($p = 6.15{\times}10^{-7}$). This demonstrates the power advantage of the pathway meta-analysis over the approach that simply combines the pathways p-values from individual studies. The aforementioned simulation studies also confirmed this observation (Table 2).

In Fig 3, we illustrate the connection between the 43 significant pathways and a group of genes showing association evidence. For the purpose of illustration, in the figure we only focus on 46 genes that are covered by the 43 pathways and have their gene-level meta-analysis p-values less than 0.001. It is evident from Fig 3 that a cluster of 4 genes, *UBE2Z*, *SNF8*, *GIP*, and *ATP5G1*, has the most significant gene-level p-values (S3 Table), and contribute association signals to 20 out of 43 significant pathways (S6–S25 Figs). These 4 genes overlap each other at chromosome 17q21. This region contains a previously unidentified genome-wide significant synonymous SNP rs1058018 (meta-analysis $p = 3.06{\times}10^{-8}$) after two rounds of inflation adjustments. More detailed information on SNP rs1058018 and SNPs in that region are given

Table 3. Summary of 43 significant pathways detected by the pathway meta-analysis based on the DIAGRAM and GERA studies.

Pathway	META[a]	DIAGRAM[b]	GERA[c]
SCHLOSSER_SERUM_RESPONSE_UP[d]	2.50E-08	2.92E-04	1.77E-03
PENG_RAPAMYCIN_RESPONSE_DN[ef]	1.50E-07	1.68E-03	2.08E-04
YAGI_AML_WITH_T_8_21_TRANSLOCATION[ef]	1.50E-07	4.33E-03	4.46E-04
PATIL_LIVER_CANCER[d]	2.00E-07	4.35E-05	3.89E-03
PUJANA_CHEK2_PCC_NETWORK[ef]	2.00E-07	1.18E-02	3.39E-03
STEIN_ESRRA_TARGETS[ef]	2.00E-07	9.37E-04	8.39E-04
STEIN_ESRRA_TARGETS_UP[ef]	3.00E-07	6.38E-03	1.02E-04
WANG_CISPLATIN_RESPONSE_AND_XPC_UP[ef]	4.00E-07	6.75E-03	1.18E-01
CADWELL_ATG16L1_TARGETS_DN[e]	4.35E-07	1.45E-03	9.59E-03
SONG_TARGETS_OF_IE86_CMV_PROTEIN[d]	5.30E-07	7.88E-04	3.20E-05
CASORELLI_ACUTE_PROMYELOCYTIC_LEUKEMIA_DN[ef]	5.50E-07	2.48E-02	5.70E-03
RIZ_ERYTHROID_DIFFERENTIATION[d]	6.15E-07	2.33E-02	2.31E-02
BORCZUK_MALIGNANT_MESOTHELIOMA_UP[d]	6.50E-07	7.61E-03	2.52E-02
HILLION_HMGA1_TARGETS[e]	9.00E-07	3.39E-01	1.10E-05
KEGG_MATURITY_ONSET_DIABETES_OF_THE_YOUNG[d]	1.14E-06	1.68E-02	3.58E-04
HOLLEMAN_ASPARAGINASE_RESISTANCE_B_ALL_DN[e]	1.22E-06	3.42E-02	1.67E-03
PUJANA_BRCA1_PCC_NETWORK[ef]	1.50E-06	1.69E-02	5.11E-03
HOSHIDA_LIVER_CANCER_SUBCLASS_S3[d]	1.95E-06	6.14E-03	5.83E-03
GRAESSMANN_APOPTOSIS_BY_DOXORUBICIN_DN[ef]	2.00E-06	9.57E-04	6.41E-04
REACTOME_REGULATION_OF_BETA_CELL_DEVELOPMENT[d]	2.26E-06	4.81E-02	1.15E-03
PUJANA_BREAST_CANCER_WITH_BRCA1_MUTATED_UP[d]	2.48E-06	7.61E-03	2.39E-02
BLALOCK_ALZHEIMERS_DISEASE_UP[d]	2.50E-06	3.52E-02	3.26E-02
GOBERT_OLIGODENDROCYTE_DIFFERENTIATION_UP[d]	2.85E-06	3.68E-02	5.37E-04
MCBRYAN_PUBERTAL_BREAST_4_5WK_DN[d]	2.95E-06	2.20E-01	1.10E-05
REACTOME_REGULATION_OF_GENE_EXPRESSION_IN_BETA_CELLS[d]	3.11E-06	2.81E-02	2.33E-03
SANSOM_APC_TARGETS_DN[ef]	3.25E-06	3.08E-01	2.84E-03
NABA_MATRISOME[d]	3.45E-06	4.46E-02	1.30E-02
PUJANA_BRCA2_PCC_NETWORK[d]	4.65E-06	1.25E-02	6.32E-02
KEGG_TYPE_II_DIABETES_MELLITUS[d]	4.85E-06	6.38E-02	9.93E-04
LINDGREN_BLADDER_CANCER_CLUSTER_1_DN[d]	5.50E-06	5.20E-02	4.36E-03
ROPERO_HDAC2_TARGETS[e]	6.04E-06	2.11E-02	1.41E-03
KEGG_INSULIN_SIGNALING_PATHWAY[d]	6.20E-06	2.65E-02	4.26E-03
CHEN_PDGF_TARGETS[d]	6.36E-06	2.48E-03	1.04E-02
REACTOME_INTEGRATION_OF_ENERGY_METABOLISM[e]	6.50E-06	6.11E-02	1.06E-04
PETROVA_ENDOTHELIUM_LYMPHATIC_VS_BLOOD_UP[d]	6.60E-06	1.79E-01	8.72E-03
REACTOME_PPARA_ACTIVATES_GENE_EXPRESSION[d]	6.70E-06	4.97E-02	1.69E-02
AGUIRRE_PANCREATIC_CANCER_COPY_NUMBER_UP[e]	6.90E-06	1.73E-01	1.97E-04
DACOSTA_UV_RESPONSE_VIA_ERCC3_UP[d]	7.45E-06	2.20E-02	4.25E-02
TOYOTA_TARGETS_OF_MIR34B_AND_MIR34C[d]	8.10E-06	1.20E-01	2.67E-03
HOLLEMAN_ASPARAGINASE_RESISTANCE_ALL_DN[e]	8.43E-06	5.90E-02	3.22E-03
REACTOME_CLASS_I_MHC_MEDIATED_ANTIGEN_PROCESSING_PRESENTATION[e]	8.45E-06	5.67E-02	2.00E-02
DODD_NASOPHARYNGEAL_CARCINOMA_DN[ef]	1.00E-05	2.86E-03	1.71E-04
REACTOME_MEMBRANE_TRAFFICKING[e]	1.04E-05	6.43E-03	4.52E-04

The 43 pathways are identified among 4,713 candidate pathways for having their pathway meta-analysis p-values less than the $<1.06\times10^{-5}$, the Bonferroni correction threshold.

[a] P-values based on summary statistics combined from the DIAGRAM and GERA studies;

[b] P-values based on summary statistics from the DIAGRAM study;

[c] P-values based on summary statistics from the GERA study;

[d] Pathways that do not contain genes in the 17q21 region;

[e] Pathways that contain at least one gene in the 17q21 region;

[f] Pathways that remain globally significant after excluding genes in the 17q21 region.

doi:10.1371/journal.pgen.1006122.t003

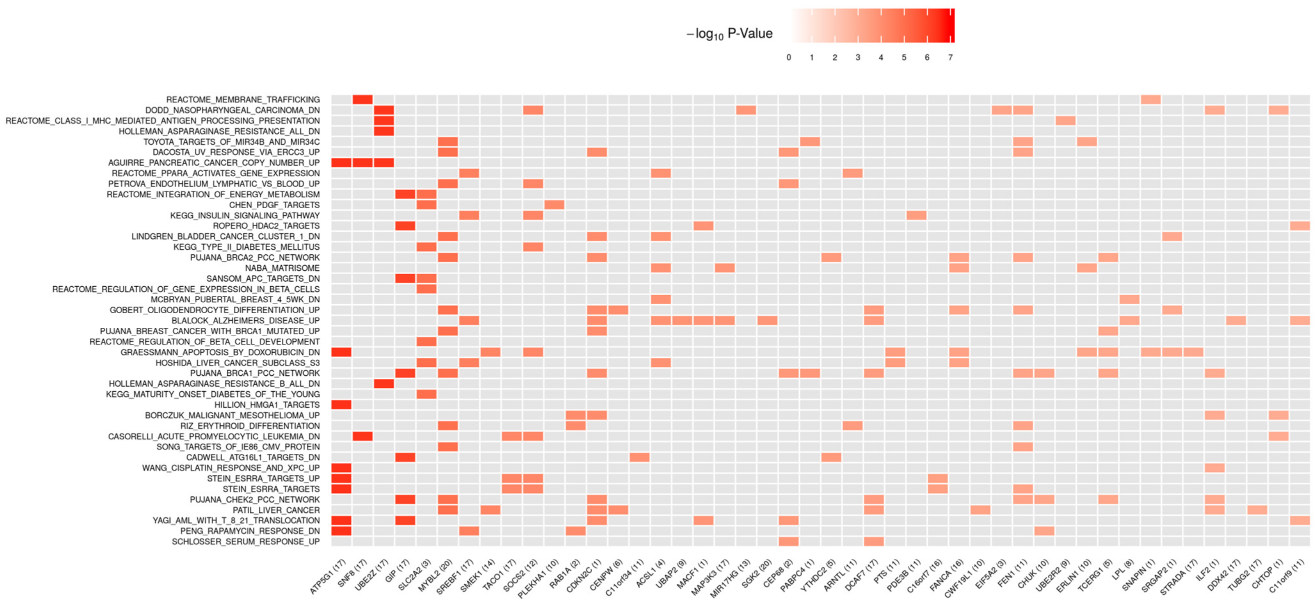

Fig 3. Heat map of gene-level p-values on selected genes within 43 significant pathways based on the DIAGRAM and GERA studies. There are 46 unique genes in the 43 significant pathways that have their gene-level meta-analysis p-values less than 0.001. Each row in the plot represents one of 43 significant pathways. Each column represents one of the 46 unique genes. The chromosome IDs of 46 unique genes are given in parentheses. The color of each cell represents the gene-level p-value (in the $-\log_{10}$ scale). A cell for a gene that is not included in a pathway is colored gray in the corresponding entry. The orders of genes (x-axis) and pathways (y-axis) are arranged according to their gene and pathway meta-analysis p-values.

doi:10.1371/journal.pgen.1006122.g003

in S4 Table, S4 and S5 Figs. By conditioning on rs1058018, none of the other SNPs in this region are significant based on the conditional association analysis using the GERA individual-level GWAS data. Based on GTEx data v6, rs1058018 is a *cis* eQTL for *UBE2Z* in blood ($p = 7.9{\times}10^{-15}$). *UBE2Z* is involved in Class I MHC antigen processing and presentation (GeneCards). The region at 17q21 was previously implicated to be associated with T2D through a candidate gene/loci approach [36]. Although genes at the 17q21 region carry the strongest association signal, 11 out of those 20 pathways remain to be globally significant ($p < 1.06{\times}10^{-5}$) after excluding those genes from the pathway definition (Table 3).

The majority of 43 identified pathways are enriched with signals from multiple chromosomal regions as demonstrated by the Q-Q plots of their SNP-level and gene-level p-values (S6–S48 Figs). For example, the strongest T2D-associated pathway, SCHLOSSER_SERUM_RESPONSE_UP, consists of 103 genes, which includes two genes with p-values < 0.001, and has 20 genes with p-values between 0.001 and 0.05 (S26 Fig, and Supplemental data). We conducted the ingenuity pathway analysis on those 22 genes with p-values less than 0.05, and found enrichment of these genes in caveolae-mediated cytosis (important for removal of low/high density lipoproteins), and lipid metabolism pathways, and in functions/diseases related to differentiation of phagocytes and transport of proteins.

It is assuring that our pathway analysis detected several pathways that are natural candidates underlying the development of T2D, including the pathways KEGG_MATURITY_ONSET_DIABETES_OF_THE_YOUNG (S31 Fig), KEGG_TYPE_II_DIABETES_MELLITUS (S41 Fig), KEGG_INSULIN_SIGNALING_PATHWAY (S43 Fig), and REACTOME_REGULATION_OF_BETA_CELL_DEVELOPMENT (S33 Fig). It is worth emphasizing that these pathways were analyzed after excluding genes in the neighborhood of 170 GWAS established T2D loci and 195 SNPs with p-values $<10^{-7}$ on either DIAGRAM or GERA data, which suggests that these well-defined T2D-related pathways are enriched with additional unidentified and contributory T2D-associated genes.

Among the 43 globally significant pathways, there are multiple ones that are defined according to specific gene expression patterns on various tumor types, including pancreatic adenocarcinoma (S21 Fig), hepatocellular carcinoma (HCC) (S27 and S32 Figs), bladder carcinoma (S42 Fig), nasopharyngeal carcinoma (S24 Fig), and familial breast cancer (S34 and S37 Figs). It is well recognized that T2D patients have elevated risk of cancer at multiple cancer sites, such as the liver and pancreas [37, 38]. These findings can provide valuable insights into the genetic basis underlying the connection between T2D and a host of different cancers.

In the above analysis, we used the sARTP method with the gene-level association evidence summarized by one or two most significant SNPs within each gene, under the assumption that there are at most two independent association signals within a given gene. We also applied sARTP by using 3 SNP-level cut points (i.e., $(c_1,c_2,c_3) = (1,2,3)$) to reanalyze the 4,713 pathways based on the combined data of DIAGRAM and GERA. It appears that results obtained by sARTP with 3 SNP-level cut points are very consistent with those with 2 cut points (S49 Fig).

Findings from Eastern Asian populations. We reanalyzed the 43 significant pathways identified from the European populations using summary-level data generated by the AGEN-T2D study. An inflation factor $\lambda_{GC} = 1.03$ calculated from the AGEN-T2D meta-analysis was adjusted in the pathway meta-analysis. The genetic regions excluded from analyzing the DIAGRAM and GERA studies were also excluded from the AGEN study. The results were summarized in Table 4. There are 10 out of 43 pathways with the unadjusted p-value less than 0.05, suggesting that many pathways identified from the European populations were also enriched with T2D-associated genes in the eastern Asian populations (S7 Table). Among the 43 pathways, we were able to identify 4 significant T2D-associated pathways at the false discovery rate (FDR [39]) of 0.05 (S27, S12, S32 and S36 Figs), and 3 additional T2D-associated pathways at the FDR of 0.1 (S47, S44, and S25 Figs). All the pathway p-values remain basically the same level if we further excluded genes within ±500kb regions surrounding the GWAS T2D loci established in eastern Asian populations. These results support the presence of trans-ethnic pathway effect on T2D in European and eastern Asian populations [11, 12].

Given the existing epidemiologic evidence on the close connection between T2D and the liver cancer, it is noteworthy that the two HCC related pathways (S27 and S32 Figs) identified in European populations remain to be significant in eastern Asian populations at the FDR of 0.05 (Table 4). The pathway PATIL_LIVER_CANCER consists of 653 genes (after data preprocessing) that are highly expressed in HCC and are enriched with genes having functions related to cell growth, cell cycle, metabolism, and cell proliferation [40]. The other pathway, HOSHIDA_LIVER_CANCER_SUBCLASS_S3 consists of 240 genes that show similar gene expression variation patterns and together define a HCC subtype with its unique histologic, molecular and clinical characteristics [41]. These two pathways have only 6 genes in common, and none of the 6 genes has a gene-level p-value < 0.05 in either European or eastern Asian data. More in depth investigations of these two complementary pathways could lead to further understanding the connection between T2D and the liver cancer.

The genome-wide significant SNP rs1058018 at the 17q21 region identified through the combined analysis of DIAGRAM and GERA studies turned out to be null in the AGEN-T2D study ($p = 0.29$). This could be due to the relatively small sample size of the AGEN-T2D study, or the genetic risk heterogeneity at the 17q21 locus among different ethnic populations. Nevertheless, 2 out of the 20 pathways (S12 and S25 Figs) that contain genes within the 17q21 region are still significant at the FDR of 0.1. Among the 23 pathways that do not contain any gene within the 17q21 region, 5 pathways remain significant at the FDR of 0.1 (S27, S32, S36, S47 and S44 Figs).

Table 4. Pathway p-values and FDR adjusted p-values based on the AGEN-T2D study.

Pathway	P-value[a]	FDR[b]
PATIL_LIVER_CANCER[c]	0.0014	0.029
CADWELL_ATG16L1_TARGETS_DN	0.0023	0.029
HOSHIDA_LIVER_CANCER_SUBCLASS_S3[c]	0.0025	0.029
GOBERT_OLIGODENDROCYTE_DIFFERENTIATION_UP[c]	0.0027	0.029
DACOSTA_UV_RESPONSE_VIA_ERCC3_UP[c]	0.011	0.074
CHEN_PDGF_TARGETS[c]	0.011	0.074
REACTOME_MEMBRANE_TRAFFICKING	0.012	0.074
AGUIRRE_PANCREATIC_CANCER_COPY_NUMBER_UP	0.026	0.14
MCBRYAN_PUBERTAL_BREAST_4_5WK_DN[c]	0.041	0.19
LINDGREN_BLADDER_CANCER_CLUSTER_1_DN[c]	0.043	0.19
PUJANA_CHEK2_PCC_NETWORK	0.057	0.21
BLALOCK_ALZHEIMERS_DISEASE_UP[c]	0.059	0.21
SCHLOSSER_SERUM_RESPONSE_UP[c]	0.085	0.27
RIZ_ERYTHROID_DIFFERENTIATION[c]	0.089	0.27
DODD_NASOPHARYNGEAL_CARCINOMA_DN	0.097	0.28
TOYOTA_TARGETS_OF_MIR34B_AND_MIR34C[c]	0.10	0.28
REACTOME_CLASS_I_MHC_MEDIATED_ANTIGEN_PROCESSING_PRESENTATION	0.14	0.32
PUJANA_BRCA2_PCC_NETWORK[c]	0.14	0.32
WANG_CISPLATIN_RESPONSE_AND_XPC_UP	0.14	0.32
STEIN_ESRRA_TARGETS	0.16	0.35
PUJANA_BREAST_CANCER_WITH_BRCA1_MUTATED_UP[c]	0.17	0.35
GRAESSMANN_APOPTOSIS_BY_DOXORUBICIN_DN	0.18	0.36
PUJANA_BRCA1_PCC_NETWORK	0.23	0.43
HOLLEMAN_ASPARAGINASE_RESISTANCE_B_ALL_DN	0.35	0.62
PENG_RAPAMYCIN_RESPONSE_DN	0.38	0.62
ROPERO_HDAC2_TARGETS	0.38	0.62
KEGG_MATURITY_ONSET_DIABETES_OF_THE_YOUNG[c]	0.39	0.62
HILLION_HMGA1_TARGETS	0.43	0.65
HOLLEMAN_ASPARAGINASE_RESISTANCE_ALL_DN	0.44	0.65
PETROVA_ENDOTHELIUM_LYMPHATIC_VS_BLOOD_UP[c]	0.45	0.65
REACTOME_PPARA_ACTIVATES_GENE_EXPRESSION[c]	0.52	0.72
REACTOME_REGULATION_OF_GENE_EXPRESSION_IN_BETA_CELLS[c]	0.55	0.74
KEGG_INSULIN_SIGNALING_PATHWAY[c]	0.58	0.74
STEIN_ESRRA_TARGETS_UP	0.59	0.74
YAGI_AML_WITH_T_8_21_TRANSLOCATION	0.64	0.76
NABA_MATRISOME[c]	0.65	0.76
KEGG_TYPE_II_DIABETES_MELLITUS[c]	0.67	0.76
SANSOM_APC_TARGETS_DN	0.69	0.76
REACTOME_INTEGRATION_OF_ENERGY_METABOLISM	0.70	0.76
REACTOME_REGULATION_OF_BETA_CELL_DEVELOPMENT[c]	0.70	0.76
BORCZUK_MALIGNANT_MESOTHELIOMA_UP[c]	0.75	0.79
SONG_TARGETS_OF_IE86_CMV_PROTEIN[c]	0.86	0.88
CASORELLI_ACUTE_PROMYELOCYTIC_LEUKEMIA_DN	0.88	0.88

These 43 pathways are nominated through the pathway meta-analysis on DIAGRAM and GERA studies. The analysis is carried out on the summary data from the AGEN-T2D study.

[a] P-values based on summary statistics from the AGEN-T2D study;

[b] FDR adjusted p-values;

[c] Pathways that do not contain genes in the 17q21 region.

doi:10.1371/journal.pgen.1006122.t004

Discussion

We developed a general statistical procedure sARTP for pathway analysis using SNP-level summary statistics generated from multiple GWAS. By applying sARTP to summary statistics from two large studies with a total of 19,809 T2D cases and 111,181 controls with European ancestry, we were able to identify 43 globally significant T2D-associated pathways after excluding genes in neighborhoods of GWAS established T2D loci. Using summary data generated from 8 T2D GWAS with 6,952 cases and 11,865 controls from eastern Asian populations, we further showed that 7 out of 43 pathways identified in the European populations were also significant in the eastern Asian populations at the FDR of 0.1. The analysis clearly highlights novel T2D-associated genes and pathways beyond what has been known from single-SNP association analysis reported from largest GWAS to date. Since the new procedure requires only SNP-level summary statistics, it provides a flexible way for conducting pathway analysis, alleviating the burden of handling large volumes of individual-level GWAS data.

We have developed a computationally efficient R package called ARTP2 implementing the ARTP and sARTP procedures, so that it can be used for conducting pathway analysis based on individual-level genetic data, as well as SNP-level summary data from one or multiple GWAS. The R package also supports the parallelization on Unix-like OS, which can substantially accelerate the computation of small p-values when a large number of resampling steps are needed. The ARTP2 package has a user-friendly interface and provides a comprehensive set of data preprocessing procedures to ensure that all the input information (e.g., allele information of SNP-level summary statistics and genotype reference panel) can be processed coherently. To make the sARTP method accessible to a wider research community, we have also developed a web-based tool that allows investigators to conduct their pathway analyses using the computing resource at the National Cancer Institutes through simple on-line inputs of summary data.

Single-locus analysis of GWAS usually has its genomic control inflation factor larger than 1.0. Some proportion of the inflation can be attributed to various confounding biases, such as the one caused by population stratification, while the other part can be due to the real polygenic effect. In the pathway analysis it is important to minimize the confounding bias at the SNP-level summary statistic. Otherwise a small bias at the SNP level can be accumulated in the pathway analysis, and lead to an elevated false discovery rate. Here we try to remove the confounding bias by adjusting for the genomic control inflation factor observed at the GWAS study. This approach is conservative because part of the inflation can be caused by the real polygenic effect. Recently, [42] developed the LD score regression method to quantify the level of inflation caused solely by the confounding bias. Adjusting for the inflation factor estimated by this method, instead of the genomic control inflation factor, can potentially increase the power of the pathway analysis. However, the LD score regression method relies on a specific polygenic risk model, and its estimate might not be robust for this model assumption. More investigations are needed to evaluate the impact of this new inflation adjustment on the pathway analysis.

There are several other strategies to increase the power of pathway analysis besides increasing sample size [4]. One area of active research is to find better ways to define the gene-level summary statistic using observed genotypes on multiple SNPs, so that it can accurately characterize the impact of the gene on the outcome [43–46]. In our proposed procedure, we adopt a data driven approach to select a subset of SNPs within a gene that collectively show the strongest association evidence. Because of this, we have to pay the penalty of multiple-comparison in the final pathway significance assessment. However, it is well recognized that SNPs at different loci can have varied levels of functional implications. We can potentially reduce the burden of multiple-comparisons and thus improve the power of the pathway analysis, by prioritizing SNPs according to existing genomic knowledge and other data resources. For example, [47] recently proposed a new gene-level summary statistic based on a prediction model that was trained with external transcriptome data. The gene-level summary statistic is defined as the predicted value that estimates the component of gene expression regulated by a subject's genotypes within the neighborhood of the considered gene. Pathway analysis procedures using this kind of biologically informed gene-level summary statistic can be easily incorporated into the ARTP2 framework.

The sARTP method can be easily expanded to adopt other multi-locus statistics in accumulating association within a gene, as long as they can be written in terms of SNP-level score statistics and their variance-covariance matrix. For example, the current ARTP2 package provides the option for conducting the pathway meta-analysis using the joint test statistics proposed by [31].

When conducting pathway analysis with individual-level genetic data, we could run into a computing memory issue if the study has a large sample size and the pathway consists of a large number of genes and SNPs (S4 Fig). The ability of performing pathway analysis using summary data provides a convenient and efficient solution in those situations. We can first calculate the SNP-level summary statistics based on the individual-level genetic data, and then randomly sample a small proportion of the original data as an internal reference to estimate the variance-covariant matrix for score statistics at considered SNPs. Based on our experiments, using 500 or more subjects to form a reference panel would be good enough to generate accurate pathway p-values. As shown in Fig 1, the testing results using this approach are very consistent with those based on individual-level genotype data.

The sARTP approach can be applied directly to SNP-level meta-analysis results. This is very convenient as meta-analysis results are in general easily accessible. But we want to emphasize that it is important to know the set of the SNPs studied by each participating study in order to apply sARTP properly, as the SNP coverage information is essential for accurately estimating the variance-covariance matrix of SNP-level score statistics. GWAS consortia usually do not post the SNP coverage information when releasing their meta-analysis results. Many statistical packages designed for conducting multi-locus analysis based on meta-analysis results often assume the uniform coverage [15–18, 24, 25, 48]. As we already have demonstrated in the context of pathway analysis, this type of over-simplification could lead to inflated false positive rate.

The proposed procedure assumes that all participating studies are conducted with subjects with the same ancestry background. If this is not the case, a simple approach is to use the Fisher's method to combine pathway p-values estimated on different ethnic populations. However, if there were no evidence for the existence of cross ethnic risk heterogeneity, it would be more powerful to assume a fixed effects model on the SNP-level association when performing the pathway analysis. In that case, since the LD structures in different ethnic populations are different, we need a separate reference panel for each ethic group to derive the corresponding variance-covariance matrix of the score statistics. The current ARTP2 package needs to be modified to accommodate such a more complicated case.

As already demonstrated by many successful GWAS meta-analysis, increasing the sample size through combining results from multiple studies is a very effective way to improve our chance for new findings. For the same reason, pathway-based meta-analysis can provide us with new opportunities to uncover biological pathways that are previously undetectable due to the limitation on the sample size. With more summary data from meta-analysis becoming increasingly available, we expect the ARTP2 package would be a valuable tool for further exploring the genome in search for the hidden heritability.

Web Resources

The URLs for data and software presented herein are as follows:

DIAbetes Genetics Replication And Meta-analysis (DIAGRAMv3), http://diagram-consortium.org/

Genetic Epidemiology Research on Aging (GERA, dbGaP Study Accession: phs000674.v1.p1), http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000674.v1.p1

Molecular Signatures Database (C2: curated gene sets), http://software.broadinstitute.org/gsea/msigdb/collections.jsp#C2

BioMart (Homo sapiens genes NCBI36 and GRCh37.p13), http://feb2014.archive.ensembl.org/

IMPUTE2, https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

GWAS Catalog, http://www.ebi.ac.uk/gwas/

1000 Genomes Project (Phase 3, v5, 2013/05/02), ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

aSPU, https://cran.r-project.org/web/packages/aSPU/index.html

GTEx Portal v6, http://gtexportal.org/home/

GeneCards Human Gene Database, http://www.genecards.org/

Ingenuity Pathway Analysis, http://www.ingenuity.com/

LocusZoom, http://locuszoom.sph.umich.edu/locuszoom/

ARTP2 package, https://cran.r-project.org/web/packages/ARTP2/

Web-based tool of ARTP2, http://analysistools.nci.nih.gov/pathway/

Supporting Information

S1 Table. Power comparison between sARTP and aSPUsPath under the scenario where each outcome-associated gene contains one functional SNP.
(DOCX)

S2 Table. Power comparison between sARTP and aSPUsPath under the scenario where each outcome-associated gene contains one or two functional SNP(s) with equal probability.
(DOCX)

S3 Table. Summary of top 50 genes with smallest gene-level p-values from the gene-level meta-analysis based on the DIAGRAM and GERA studies.
(DOCX)

S4 Table. Effect of SNP rs1058018 on type 2 diabetes.
(DOCX)

S5 Table. The genomic control inflation factors, Spearman's rank correlation coefficients between the pathway size and its p-value based on results obtained by applying sARTP to 20 simulated GWAS under the null.
(DOCX)

S6 Table. Gene-level p-values of genes in the 43 significant pathways identified in the European populations based on combined data from the DIAGRAM and GERA studies.
(XLSX)

S7 Table. Gene-level p-values of genes in the 10 pathways with unadjusted pathway p-values < 0.05 based on data from the AGEN-T2D study.
(XLSX)

S1 Text. Recovering score statistics and their variance-covariance matrix using summary results from the fixed effects model.
(DOCX)

S2 Text. Further evaluation of sARTP under the null.
(DOCX)

S1 Fig. The sample size used for each SNP in the DIAGRAM meta-analysis. The SNP index (x-axis) is sorted according to its meta-analysis sample size. There are 2.5 million autosomal SNPs genotyped or imputed in at least one of the twelve participating GWAS in the DIAGRAM meta-analysis. 19.0% of those SNPs have a sample size of 69,033 (12,171 cases and 56,862 controls), which is total sample size of all twelve participating GWAS combined. Another 57.9% of those SNPs have a sample size of 63,390 (9,580 cases and 53,810).
(TIF)

S2 Fig. Histograms of numbers of SNPs and genes after SNP filtering within each of 4,718 pathways in pathway analyses of the DIAGRAM study, the GERA study, and the two studies combined (META).
(TIF)

S3 Fig. Boxplot of the number of pathways containing genes with p-values in a given range. Genes are stratified into five groups according to their p-values from the gene-level meta-analysis on the summary data from the DIAGRAM and GERA studies. The boxplot summarizes the number of pathways containing a gene within a given group.
(TIF)

S4 Fig. The LocusZoom plot showing ±100kb region of rs1058018 in European populations. The SNP p-values were computed based on combined data of DIAGRAM and GERA studies after two rounds of genomic control inflation factor adjustment.
(TIF)

S5 Fig. The LocusZoom plot showing ±100kb region of rs1058018 in eastern Asian populations. The SNP p-values were computed based on data of AGEN-T2D study after the adjustment of genomic control inflation factor.
(TIF)

S6 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway PENG_RAPAMYCIN_RESPONSE_DN.
(TIF)

S7 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway YAGI_AML_WITH_T_8_21_TRANSLOCATION.
(TIF)

S8 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway PUJANA_CHEK2_PCC_NETWORK.
(TIF)

S9 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway STEIN_ESRRA_TARGETS.
(TIF)

S10 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway STEIN_ESRRA_TARGETS_UP.
(TIF)

S11 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway WANG_CISPLATIN_RESPONSE_AND_XPC_UP.
(TIF)

S12 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway CADWELL_ATG16L1_TARGETS_DN.
(TIF)

S13 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway CASORELLI_ACUTE_PROMYELOCYTIC_LEUKEMIA_DN.
(TIF)

S14 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway HILLION_HMGA1_TARGETS.
(TIF)

S15 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway HOLLEMAN_ASPARAGINASE_RESISTANCE_B_ALL_DN.
(TIF)

S16 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway PUJANA_BRCA1_PCC_NETWORK.
(TIF)

S17 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway GRAESSMANN_APOPTOSIS_BY_DOXORUBICIN_DN.
(TIF)

S18 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway SANSOM_APC_TARGETS_DN.
(TIF)

S19 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway ROPERO_HDAC2_TARGETS.
(TIF)

S20 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway REACTOME_INTEGRATION_OF_ENERGY_METABOLISM.
(TIF)

S21 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway AGUIRRE_PANCREATIC_CANCER_COPY_NUMBER_UP.
(TIF)

S22 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway HOLLEMAN_ASPARAGINASE_RESISTANCE_ALL_DN.
(TIF)

S23 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway REACTOME_CLASS_I_MHC_MEDIATED_ANTIGEN_PROCESSING_PRESENTATION.
(TIF)

S24 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway DODD_NASOPHARYNGEAL_CARCINOMA_DN.
(TIF)

S25 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway REACTOME_MEMBRANE_TRAFFICKING.
(TIF)

S26 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway SCHLOSSER_SERUM_RESPONSE_UP.
(TIF)

S27 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway PATIL_LIVER_CANCER.
(TIF)

S28 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway SONG_TARGETS_OF_IE86_CMV_PROTEIN.
(TIF)

S29 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway RIZ_ERYTHROID_DIFFERENTIATION.
(TIF)

S30 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway BORCZUK_MALIGNANT_MESOTHELIOMA_UP.
(TIF)

S31 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway KEGG_MATURITY_ONSET_DIABETES_OF_THE_YOUNG.
(TIF)

S32 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway HOSHIDA_LIVER_CANCER_SUBCLASS_S3.
(TIF)

S33 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway REACTOME_REGULATION_OF_BETA_CELL_DEVELOPMENT.
(TIF)

S34 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway PUJANA_BREAST_CANCER_WITH_BRCA1_MUTATED_UP.
(TIF)

S35 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway BLALOCK_ALZHEIMERS_DISEASE_UP.
(TIF)

S36 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway GOBERT_OLIGODENDROCYTE_DIFFERENTIATION_UP.
(TIF)

S37 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway MCBRYAN_PUBERTAL_BREAST_4_5WK_DN.
(TIF)

S38 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway REACTOME_REGULATION_OF_GENE_EXPRESSION_IN_BETA_CELLS.
(TIF)

S39 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway NABA_MATRISOME.
(TIF)

S40 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway PUJANA_BRCA2_PCC_NETWORK.
(TIF)

S41 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway KEGG_TYPE_II_DIABETES_MELLITUS.
(TIF)

S42 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway LINDGREN_BLADDER_CANCER_CLUSTER_1_DN.
(TIF)

S43 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway KEGG_INSULIN_SIGNALING_PATHWAY.
(TIF)

S44 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway CHEN_PDGF_TARGETS.
(TIF)

S45 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway PETROVA_ENDOTHELIUM_LYMPHATIC_VS_BLOOD_UP.
(TIF)

S46 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway REACTOME_PPARA_ACTIVATES_GENE_EXPRESSION.
(TIF)

S47 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway DACOSTA_UV_RESPONSE_VIA_ERCC3_UP.
(TIF)

S48 Fig. Q-Q plots for SNP p-values and sARTP gene p-values of pathway TOYOTA_TARGETS_OF_MIR34B_AND_MIR34C.
(TIF)

S49 Fig. Comparison of sARTP p-values obtained with 2 or 3 SNP-level cut points based on combined data of DIAGRAM and GERA. The p-values (in $-\log_{10}$ scale) of 4,713 pathways defined in MSigDB v5.0 were obtained with sARTP.
(TIF)

S50 Fig. Q-Q plots of pathway p-values based on 20 GWAS datasets generated under the null. Based on each generated GWAS, 4,439 pathways (each with no more than 10,000 SNPs) defined in MSigDB v5.0 were analyzed with sARTP. λ is the genomic control inflation factor of the pathway p-values.
(TIF)

Acknowledgments

This study utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov/). The authors thank the AGEN-T2D consortium and DIAGRAM consortium for sharing the meta-analysis summary data. The authors acknowledge Sue Pan, JJ Pan, Wesley Obenshain, Tony Hall, Jim Zhou, Ye Wu, Cuong Nguyen for their help in developing the web-based tool for ARTP2.

Author Contributions

Conceived and designed the experiments: KY HZ NC. Performed the experiments: HZ WW KY. Analyzed the data: HZ WW KY. Contributed reagents/materials/analysis tools: HZ WW YY JS. Wrote the paper: KY HZ PLH JS NC.

References

1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research. 2014; 42(Database issue):D1001–6. doi: 10.1093/nar/gkt1229 PMID: 24316577
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–53. doi: 10.1038/nature08494 PMID: 19812666
3. Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. Bioinformatics. 2010; 26(4):445–55. doi: 10.1093/bioinformatics/btp713 PMID: 20053841
4. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nature reviews Genetics. 2010; 11(12):843–54. doi: 10.1038/nrg2884 PMID: 21085203
5. Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, et al. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. American journal of human genetics. 2010; 86(6):860–71. doi: 10.1016/j.ajhg.2010.04.014 PMID: 20560206
6. Evangelou M, Rendon A, Ouwehand WH, Wernisch L, Dudbridge F. Comparison of methods for competitive tests of pathway analysis. PloS one. 2012; 7(7):e41018. doi: 10.1371/journal.pone.0041018 PMID: 22859961
7. Li MX, Kwan JS, Sham PC. HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. American journal of human genetics. 2012; 91(3):478–88. doi: 10.1016/j.ajhg.2012.08.004 PMID: 22958900
8. Pan W, Kwak IY, Wei P. A Powerful Pathway-Based Adaptive Test for Genetic Association with Common or Rare Variants. American journal of human genetics. 2015; 97(1):86–98. doi: 10.1016/j.ajhg.2015.05.018 PMID: 26119817
9. Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, et al. Pathway analysis by adaptive combination of P-values. Genetic epidemiology. 2009; 33(8):700–9. doi: 10.1002/gepi.20422 PMID: 19333968
10. Cho YS, Chen CH, Hu C, Long J, Ong RT, Sim X, et al. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nature genetics. 2012; 44(1):67–72.
11. DIAbetes Genetics Replication Meta-analysis C, Consortium AGENTD, Consortium SATD, Consortium MATD, Consortium TDGEbN-gsim-ES, Mahajan A, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nature genetics. 2014; 46(3):234–44. doi: 10.1038/ng.2897 PMID: 24509480
12. Imamura M, Takahashi A, Yamauchi T, Hara K, Yasuda K, Grarup N, et al. Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. Nature communications. 2016; 7:10531. doi: 10.1038/ncomms10531 PMID: 26818947
13. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics. 2012; 44(9):981–90. doi: 10.1038/ng.2383 PMID: 22885922
14. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature genetics. 2010; 42(11):937–48. doi: 10.1038/ng.686 PMID: 20935630

15. Burren OS, Guo H, Wallace C. VSEAMS: a pipeline for variant set enrichment analysis using summary GWAS data identifies IKZF3, BATF and ESRRA as key transcription factors in type 1 diabetes. Bioinformatics. 2014; 30(23):3342–8. doi: 10.1093/bioinformatics/btu571 PMID: 25170024

16. Evangelou M, Smyth DJ, Fortune MD, Burren OS, Walker NM, Guo H, et al. A method for gene-based pathway analysis using genomewide association study summary statistics reveals nine new type 1 diabetes associations. Genetic epidemiology. 2014; 38(8):661–70. doi: 10.1002/gepi.21853 PMID: 25371288

17. Kwak IY, Pan W. Adaptive gene- and pathway-trait association testing with GWAS summary statistics. Bioinformatics. 2015.

18. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLoS computational biology. 2016; 12(1): e1004714. doi: 10.1371/journal.pcbi.1004714 PMID: 26808494

19. Network and Pathway Analysis Subgroup of Psychiatric Genomics C. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. Nature neuroscience. 2015; 18(2):199–209. doi: 10.1038/nn.3922 PMID: 25599223

20. Segre AV, Consortium D, investigators M, Groop L, Mootha VK, Daly MJ, et al. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS genetics. 2010; 6(8).

21. Swanson DM, Blacker D, Alchawa T, Ludwig KU, Mangold E, Lange C. Properties of permutation-based gene tests and controlling type 1 error using a summary statistic based gene test. BMC genetics. 2013; 14:108. doi: 10.1186/1471-2156-14-108 PMID: 24199751

22. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–73. doi: 10.1038/nature09534 PMID: 20981092

23. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. doi: 10.1038/nature11632 PMID: 23128226

24. Hu YJ, Berndt SI, Gustafsson S, Ganna A, Genetic Investigation of ATC, Hirschhorn J, et al. Meta-analysis of gene-level associations for rare variants based on single-variant statistics. American journal of human genetics. 2013; 93(2):236–48. doi: 10.1016/j.ajhg.2013.06.011 PMID: 23891470

25. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A versatile gene-based test for genome-wide association studies. American journal of human genetics. 2010; 87(1):139–45. doi: 10.1016/j.ajhg.2010.06.009 PMID: 20598278

26. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ATC, Replication DIG, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature genetics. 2012; 44(4):369–75, S1–3. doi: 10.1038/ng.2213 PMID: 22426310

27. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007; 23(8):980–7. PMID: 17303618

28. Seaman SR, Muller-Myhsok B. Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. American journal of human genetics. 2005; 76(3):399–408. PMID: 15645388

29. Hoffmann TJ, Kvale MN, Hesselson SE, Zhan Y, Aquino C, Cao Y, et al. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics. 2011; 98(2):79–89. doi: 10.1016/j.ygeno.2011.04.005 PMID: 21565264

30. Hoffmann TJ, Zhan Y, Kvale MN, Hesselson SE, Gollub J, Iribarren C, et al. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics. 2011; 98 (6):422–30. doi: 10.1016/j.ygeno.2011.08.007 PMID: 21903159

31. Zhang H, Shi J, Liang F, Wheeler W, Stolzenberg-Solomon R, Yu K. A fast multilocus test with adaptive SNP selection for large-scale genetic-association studies. European Journal of Human Genetics. 2014; 22(5):696–702. doi: 10.1038/ejhg.2013.201 PMID: 24022295

32. Ge Y, Dudoit S, Speed T. Resampling-based multiple testing for microarray data analysis. Test. 2003; 12:1–77.

33. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theoretical population biology. 2001; 60(3):155–66. PMID: 11855950

34. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature reviews Genetics. 2010; 11(7):499–511. doi: 10.1038/nrg2796 PMID: 20517342

35. Wang T, Elston RC. Improved power by use of a weighted score test for linkage disequilibrium mapping. American journal of human genetics. 2007; 80(2):353–60. PMID: 17236140

36. Johnson ME, Zhao J, Schug J, Deliard S, Xia Q, Guy VC, et al. Two novel type 2 diabetes loci revealed through integration of TCF7L2 DNA occupancy and SNP association data. BMJ open diabetes research & care. 2014; 2(1):e000052.

37. Lin CC, Chiang JH, Li CI, Liu CS, Lin WY, Hsieh TF, et al. Cancer risks among patients with type 2 diabetes: a 10-year follow-up study of a nationwide population-based cohort in Taiwan. BMC cancer. 2014; 14:381. doi: 10.1186/1471-2407-14-381 PMID: 24884617

38. Tsilidis KK, Kasimis JC, Lopez DS, Ntzani EE, Ioannidis JP. Type 2 diabetes and cancer: umbrella review of meta-analyses of observational studies. Bmj. 2015; 350:g7607. doi: 10.1136/bmj.g7607 PMID: 25555821

39. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate—a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met. 1995; 57(1):289–300.

40. Patil MA, Chua MS, Pan KH, Lin R, Lih CJ, Cheung ST, et al. An integrated data analysis approach to characterize genes highly expressed in hepatocellular carcinoma. Oncogene. 2005; 24(23):3737–47. PMID: 15735714

41. Hoshida Y, Nijman SM, Kobayashi M, Chan JA, Brunet JP, Chiang DY, et al. Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. Cancer research. 2009; 69(18):7385–92. doi: 10.1158/0008-5472.CAN-09-1089 PMID: 19723656

42. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics. 2015; 47(3):291–5. doi: 10.1038/ng.3211 PMID: 25642630

43. Li M, Wang K, Grant SF, Hakonarson H, Li C. ATOM: a powerful gene-based association test by combining optimally weighted markers. Bioinformatics. 2009; 25(4):497–503. doi: 10.1093/bioinformatics/btn641 PMID: 19074959

44. Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. American journal of human genetics. 2006; 79(5):792–806. PMID: 17033957

45. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. American journal of human genetics. 2010; 86(6):929–42. doi: 10.1016/j.ajhg.2010.05.002 PMID: 20560208

46. Zhang H, Wheeler W, Wang Z, Taylor PR, Yu K. A fast and powerful tree-based association test for detecting complex joint effects in case-control studies. Bioinformatics. 2014; 30(15):2171–8. doi: 10.1093/bioinformatics/btu186 PMID: 24794927

47. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature genetics. 2015; 47(9):1091–8. doi: 10.1038/ng.3367 PMID: 26258848

48. Mishra A, Macgregor S. VEGAS2: Software for More Flexible Gene-Based Testing. Twin research and human genetics: the official journal of the International Society for Twin Studies. 2015; 18(1):86–91.