Gene and Network Analysis of Common Variants Reveals Novel Associations in Multiple Complex Diseases

Priyanka Nakka,*,† Benjamin J. Raphael,†,‡,1 and Sohini Ramachandran*,†,1
*Department of Ecology and Evolutionary Biology, †Center for Computational Molecular Biology, and ‡Department of Computer Science, Brown University, Providence, Rhode Island 02912

ABSTRACT Genome-wide association (GWA) studies typically lack power to detect genotypes significantly associated with complex diseases, where different causal mutations of small effect may be present across cases. A common, tractable approach for identifying genomic elements associated with complex traits is to evaluate combinations of variants in known pathways or gene sets with shared biological function. Such gene-set analyses require the computation of gene-level *P*-values or gene scores; these gene scores are also useful when generating hypotheses for experimental validation. However, commonly used methods for generating GWA gene scores are computationally inefficient, biased by gene length, imprecise, or have low true positive rate (TPR) at low false positive rates (FPR), leading to erroneous hypotheses for functional validation. Here we introduce a new method, PEGASUS, for analytically calculating gene scores. PEGASUS produces gene scores with as much as 10 orders of magnitude higher numerical precision than competing methods. In simulation, PEGASUS outperforms existing methods, achieving up to 30% higher TPR when the FPR is fixed at 1%. We use gene scores from PEGASUS as input to HotNet2 to identify networks of interacting genes associated with multiple complex diseases and traits; this is the first application of HotNet2 to common variation. In ulcerative colitis and waist–hip ratio, we discover networks that include genes previously associated with these phenotypes, as well as novel candidate genes. In contrast, existing methods fail to identify these networks. We also identify networks for attention-deficit/hyperactivity disorder, in which GWA studies have yet to identify any significant SNPs.

KEYWORDS GWAS; common variants; complex diseases; pathway analysis; quantitative traits

GENOME-WIDE association (GWA) studies and meta-analyses are widely used to identify susceptibility loci for complex diseases and traits, which are phenotypes generated by multiple mutations of moderate to small effect (Hirschhorn and Daly 2005; McCarthy *et al.* 2008; Daly 2010; Jiang *et al.* 2012; Evangelou and Ioannidis 2013; Nalls *et al.* 2014; Skibola *et al.* 2014; Woo *et al.* 2014; Buch *et al.* 2015; Kouri *et al.* 2015; Litchfield *et al.* 2015; Renton *et al.* 2015; Hallberg *et al.* 2016). To date, >2400 GWA studies have been conducted to find causal variants that are statistically associated with a disease or trait (http://www.ebi.ac.uk/gwas/). The GWA framework tests the hypothesis that individual mutations of large effect generate phenotypes of interest. However, this framework has multiple limitations when applied to complex diseases. First, complex diseases are known to exhibit genetic heterogeneity on multiple levels: (i) The disease may be generated by multiple mutations within an associated gene and (ii) mutations in distinct genes within a pathway may interact and produce the disease state (McClellan and King 2010). In both cases, separately testing individual variants for statistical associations with a phenotype may not identify susceptibility loci (McClellan and King 2010; Stranger *et al.* 2011). Further, SNP-level GWA results are unlikely to reveal complex disease mechanisms, given that different combinations of functionally related variants in genes and pathways may interact to produce the phenotype of interest.

Gene-set analyses, which test for the statistical association of phenotype state with a set of genes, are commonly used to address these limitations of the GWA framework (see Wang

doi: 10.1534/genetics.116.188391
Manuscript received February 25, 2016; accepted for publication July 24, 2016; published Early Online August 3, 2016.
Available freely online through the author-supported open access option.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.188391/-/DC1.
[1]Corresponding authors: Computer Science Department, Princeton University, 35 Olden Street, Princeton NJ 08540. Email: braphael@princeton.edu; and Box G-W, Brown University, Providence, RI 02912. E-mail: sramachandran@brown.edu

et al. 2010, Leiserson *et al.* 2013, and Mooney *et al.* 2014 for reviews). To increase computational efficiency and limit the number of hypotheses tested, it is necessary to reduce the combinations of variants examined to a tractable number. This is typically done using databases of known pathways or other biological interactions, nearly all of which are annotated at the gene level. Thus, a crucial step in most gene-set analyses is combining SNP-level GWA *P*-values within genes into a "gene score" (Mooney *et al.* 2014). Here we use "gene-set analysis" to describe three types of statistical tests for association at the gene level with a phenotype of interest. First, we describe permutation tests (*e.g.*, DAPPLE and dmGWAS) (Jia *et al.* 2011; Rossin *et al.* 2011) where *P*-values are assigned to gene scores observed in an annotated pathway. Second, we describe tests for enrichment in related annotations among genes in a predetermined list (*e.g.*, GRAIL, MAGENTA, DAVID, and GSEA-SNP) (Huang *et al.* 2007; Holden *et al.* 2008; Raychaudhuri *et al.* 2009; Segrè *et al.* 2010). To conduct these tests, the investigator must compute a gene score and, in some cases, determine a threshold for extreme gene scores to generate a list of genes associated with the phenotype of interest. In the third type of test, once a gene score is computed, the investigator can conduct a gene-level association test and/or a gene-network association test, to identify novel combinations of variants that generate a phenotype of interest, due to unknown interactions between genes or uncharacterized crosstalk between pathways.

An informative gene score is a necessary ingredient for accurate gene-set analyses, but all commonly used methods for generating gene scores have substantial drawbacks. Commonly used methods include choosing the best SNP *P*-value within a gene to be the gene score, which is sometimes referred to as "minSNP" (Torkamani *et al.* 2008; Fehringer *et al.* 2012; Gelernter *et al.* 2015; Hu *et al.* 2015); permutation-based methods such as permSNP (Wang *et al.* 2007; Eleftherohorinou *et al.* 2009; Ballard *et al.* 2010; Christoforou *et al.* 2014; Evangelou *et al.* 2014; Backes *et al.* 2016); regression-based methods such as the sequence kernel association test (SKAT) family of tests (Wu *et al.* 2010, 2011) and stratified LD score regression (Finucane *et al.* 2015); and VEGAS (Liu *et al.* 2010) and RAREMETALS (Liu *et al.* 2013), which correct for linkage disequilibrium (LD) between SNPs, using simulations from a multivariate normal distribution whose variance is the empirical LD observed among SNPs within each gene being analyzed. Multiple methods exist that use the same null distribution as VEGAS (Tzeng and Zhang 2007; Pan 2009). Other methods that have been proposed include Fisher's combination test (where the gene score must be calculated empirically using permutation tests), Simes' combination test, and Sidak's combination test (Ballard *et al.* 2010; Peng *et al.* 2010; Wojcik *et al.* 2015).

The limitations of these approaches range from biased results to computational inefficiency to imprecision (Table 1). minSNP is heavily biased by gene length; the longer the gene is, the more likely it is to have a low gene score. permSNP permutes case–control labels within a genotype data set to calculate an empirical *P*-value for every gene, which becomes computationally intractable for large data sets (Liu *et al.* 2010). Further, permSNP and SKAT require gaining access to genotype data to perform permutations and regression, respectively. The VEGAS method is more computationally efficient than other permutation methods (*e.g.*, permSNP requires recomputing GWA *P*-values for each permuted data set) and requires only GWA SNP-level *P*-values as input, but both permSNP and VEGAS give gene scores whose numerical precision depends on the number of permutations and simulations, respectively, that are performed. The smallest gene score VEGAS reports by default is 10^{-6} and for permSNP, it is the reciprocal of the number of permutations performed per gene (Table 1).

Here we propose a new method—the precise, efficient gene association score using SNPs (PEGASUS)—to calculate gene scores analytically from a null chi-square distribution that captures LD between SNPs in a gene and addresses the shortcomings of existing methods. PEGASUS requires only GWA study summary statistics and a suitable reference population for LD calculations as input and thus can be applied to GWA study meta-analyses performed on summary statistics, pooled DNA sequencing GWA studies, family-based GWA studies, transmission disequilibrium test (TDT) results, and also traditional GWA studies where consent guidelines prohibit release of genotype data. PEGASUS gene scores are correlated with statistics like VEGAS (Tzeng and Zhang 2007; Pan 2009; Liu *et al.* 2010), which rely on the same null distributions to calculate gene scores. These methods use different approximations for the distribution of the sum of correlated chi-square statistics, in contrast to the more accurate numerical integration of the null distribution implemented in PEGASUS. We apply our method to publicly available GWA data sets for nine common diseases and three quantitative traits from the Psychiatric Genomics Consortium (PGC) (Neale *et al.* 2010; Ripke *et al.* 2011, 2013; Sklar *et al.* 2011), the International IBD Genetics Consortium (IIBDGC) (Franke *et al.* 2010; Anderson *et al.* 2011), the Genetic Investigation of Anthropometric Traits (GIANT) Consortium (Heid *et al.* 2010; Lango Allen *et al.* 2010; Speliotes *et al.* 2010), the Broad Institute (Stahl *et al.* 2010), the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium (Morris *et al.* 2012), and Xu *et al.* (2013) (Table 2). These data sets were chosen because the full set of SNP-level *P*-values from the GWA study were available for public download. We compare our method to gene scores generated by minSNP, permSNP, SKAT, and VEGAS, using real and simulated GWA data. Finally, we use our gene scores as input in pathway analysis with HotNet2 (Leiserson *et al.* 2015), thereby conducting the first application of HotNet2 to common genetic variation and identifying gene networks harboring several variants associated with three phenotypes: attention-deficit/hyperactivity disorder, ulcerative colitis, and waist–hip ratio. For these three phenotypes of interest, HotNet2 using VEGAS gene scores recovered fewer significant subnetworks for attention-deficit/hyperactivity disorder, ulcerative colitis, and waist–hip ratio. Neither VEGAS nor PEGASUS yielded significant subnetworks for the other nine traits studied here.

Table 1 Summary of minSNP, permSNP, SKAT, and VEGAS gene score methods and limitations

Gene score method	How it works	Limitations
minSNP	The gene score for each gene is the smallest SNP *P*-value observed within that gene in a GWA study.	Biased by gene length (longer genes have lower gene scores).
permSNP	Permutes case–control labels within a genotype data set, recomputes GWA SNP *P*-values using a permuted data set, and calculates an empirical gene *P*-value based on the number of times the observed average SNP *P*-value is lower than the permuted *P*-values	Requires access to genotype data; very computationally costly for genome-wide data sets; numerical precision of gene scores is bounded by the number of permutations performed.
SKAT	Uses multiple linear/logistic mixed-model regression of covariates (such as principal components to control for population stratification) and genotypes for variants in a gene set onto disease state.	Requires access to genotype data.
VEGAS	Uses simulations from a multivariate normal distribution to correct for LD between SNPs. The variance of the distribution is the empirical LD observed among SNPs within each gene in the data set.	Numerical precision of gene scores is bounded by the number of simulations performed; computationally inefficient due to simulations.

Materials and Methods

Data sets analyzed from genome-wide association studies

Methods for computing gene scores require a full list of GWA SNP-level *P*-values; these can be computed from genotype data obtained from previously published GWA studies. We were able to obtain complete results from previous GWA studies for nine common diseases and three quantitative traits (Table 2). See Supplemental Material, Table S3 for URLs to download GWA *P*-values from the studies referenced in Table 2.

Gene scores

We compared PEGASUS to four existing methods for generating gene-based scores from GWA SNP-level *P*-values: (i) minSNP (Torkamani *et al.* 2008; Fehringer *et al.* 2012; Gelernter *et al.* 2015; Hu *et al.* 2015), (ii) permSNP (Wang *et al.* 2007; Eleftherohorinou *et al.* 2009; Ballard *et al.* 2010; Christoforou *et al.* 2014; Evangelou *et al.* 2014; Backes *et al.* 2016), (iii) SKAT (Wu *et al.* 2010, 2011), and (iv) VEGAS (Liu *et al.* 2010). For n markers in a given gene, these methods use different strategies, each detailed below and summarized in Figure 1, to combine *P*-values $p_1 \ldots p_n$ within the gene and calculate a gene-level *P*-value. We refer to this gene-level *P*-value as the gene score or p_g.

minSNP: The minSNP method (Torkamani *et al.* 2008; Fehringer *et al.* 2012; Gelernter *et al.* 2015; Hu *et al.* 2015) for generating gene scores assigns the smallest GWA SNP-level *P*-value in a given gene to be the gene score (Equation 1):

$$p_g = \min(p_1, p_2, \ldots, p_n). \quad (1)$$

permSNP: The permSNP method (Wang *et al.* 2007; Eleftherohorinou *et al.* 2009; Ballard *et al.* 2010; Christo-Forou *et al.* 2014; Evangelou *et al.* 2014; Backes *et al.* 2016) produces gene scores by permuting phenotype labels across all genotyped individuals to generate an empirical *P*-value for every gene. We carried out permSNP only on the acute lymphoblastic leukemia (ALL) data set (Xu *et al.* 2013), as genotype data are required for this method, and we did not have genotype data for the other traits analyzed here. We calculated permSNP gene scores only for the top 400 most significant genes determined by minSNP using set-based test analysis in PLINK due to computational constraints (Purcell *et al.* 2007) (see File S1, Algorithm S1 for more details.)

The following settings were used to calculate permSNP gene scores in PLINK (Purcell *et al.* 2007): --set-r2 1, --set-p 1, --set-max 99999, --maf 0.01, and --mperm 10,000 permutations of case–control labels. With these command flags, PLINK first does an association test between phenotype state and allele dosage at each SNP. Second, for every gene, the SNP test statistics $(q_1, q_2, \ldots, q_n)$ within the gene are averaged to calculate the observed gene-level test statistic Q_{obs} (File S1, Algorithm S1). Third, the phenotype labels are permuted M times and the previous two steps are repeated for the permuted data each time, resulting in SNP statistics using the permuted phenotype data and corresponding gene statistics Q^*. The gene score p_g is then the fraction of times the gene statistic Q^* is greater than the observed statistic Q_{obs} over the M permutations (File S1, Algorithm S1).

SKAT (Wu et al. 2010, 2011): This method uses multiple linear/logistic mixed-model regression of covariates and genotypes for variants in a gene set, along with covariates, onto disease state. Covariates can include sex, age, or top principal components of genotype data to control for population stratification. Under the multiple logistic regression model for a continuous phenotype, the relationship between variant genotypes $\mathbf{G_i}$ and the phenotype y_i for the ith individual (of p total individuals) is given by Equation 2, where α_0 is an intercept term, $\mathbf{C_i}$ is a vector of covariates, $\boldsymbol{\alpha}$ is the vector of regression coefficients for m covariates, $\boldsymbol{\beta}$ is the vector of regression coefficients for the n SNPs in a gene, and $\boldsymbol{\epsilon}_i$ is an error term that is normally distributed with mean of zero and variance σ^2. Given this model, SKAT tests the null hypothesis

Table 2 Total numbers of cases and controls and number of SNP loci in GWA studies for the 12 phenotypes studied here

Disease or trait (reference)	No. cases	No. controls	No. SNPs
Attention-deficit/hyperactivity disorder (ADHD) (Neale *et al.* 2010)	864 + 2,064 trios	2,455	1,206,461
Acute lymphoblastic leukemia (ALL) (Xu *et al.* 2013)	1,593	6,661	709,059
Bipolar disorder (BIP) (Sklar *et al.* 2011)	7,481	9,250	2,427,220
Body mass index (BMI) (Speliotes *et al.* 2010)	NA	123,865	2,471,516
Crohn's disease (CD) (Franke *et al.* 2010)	6,333	15,056	953,241
Height (Lango Allen *et al.* 2010)	NA	183,727	2,469,635
Major depressive disorder (MDD) (Ripke *et al.* 2013)	9,240	9,519	1,235,109
Rheumatoid arthritis (RA) (Stahl *et al.* 2010)	5,539	20,169	2,556,271
Schizophrenia (SCZ) (Ripke *et al.* 2011)	9,394	12,462	1,252,901
Type 2 diabetes (T2D) (Morris *et al.* 2012)	12,171	56,862	2,473,441
Ulcerative colitis (UC) (Anderson *et al.* 2011)	6,687	19,718	1,428,749
Waist–hip ratio adjusted for BMI (WHR) (Heid *et al.* 2010)	NA	77,167	2,483,325

$H_0: \boldsymbol{\beta} = \mathbf{0}$. Assuming that each β_j for the jth variant follows some distribution with mean 0 and variance $w_j\tau$, where τ is a variance component and w_j is a weight for variant j, the null hypothesis can be restated as $H_0: \tau = 0$. The variance component score statistic for this test is given by Equation 3, where $\mathbf{K}$ is a $p \times p$ matrix with $k_{ii'} = \sum_{j=1}^{n} w_j G_{ij} G_{i'j}$, the weighted genetic similarity between two subjects i and i' in the region with n markers, and $\hat{\boldsymbol{\mu}} = \widehat{\alpha_0} + \mathbf{C}\hat{\boldsymbol{\alpha}}$. Wu *et al.* (2011) suggest setting the weights $\sqrt{w_j} = \text{Beta}(\text{MAF}_j; 1, 25)$, the beta distribution density function with parameters $a_1 = 1$ and $a_2 = 25$ evaluated at the sample minor-allele frequency (MAF) for a given variant j. The SKAT test statistic follows a mixture of chi-square distributions that can be evaluated using numerical integration to obtain a P-value for the gene (Wu *et al.* 2010, 2011):

$$y_i = \alpha_0 + \boldsymbol{\alpha}'\mathbf{C_i} + \boldsymbol{\beta}'\mathbf{G_i} + \epsilon_i \tag{2}$$

$$Q = (\mathbf{y} - \hat{\boldsymbol{\mu}})'\mathbf{K}(\mathbf{y} - \hat{\boldsymbol{\mu}}). \tag{3}$$

Because full genotype data are required for this method, we applied SKAT only on the ALL data set (Xu *et al.* 2013) and the Wellcome Trust Case Control Consortium (WTCCC) type 2 diabetes data set (WTCCC 2007). We used the top four principal components from principal components analysis (PCA) on these data sets as covariates in the regression and hold these covariates constant across all methods tested in this study (Wu *et al.* 2010, 2011; Peloso *et al.* 2014). The following settings were used to calculate SKAT scores in R, using the R package SKAT (Lee *et al.* 2015): (a) in the R function SKAT_Null_Model, out_type = "D" and Adjustment = "F" and (b) in the R function SKATBinary.SSD.All, method = "SKAT". These settings specify a linear weighted kernel with the weights $\sqrt{w_j} = \text{Beta}(\text{MAF}_j; 1, 25)$.

VEGAS (Liu et al. 2010): Consider a gene with n SNPs. Under the null hypothesis of no association, a gene can be represented by an n-element multivariate normal vector with mean 0 and variance $\sum$, the $n \times n$ pairwise LD matrix. Given this model, VEGAS generates gene scores by (i) performing 10^6 multivariate normal simulations from the null distribution of LD-correlated SNPs, (ii) squaring the simulated values and summing to get a null test statistic for each gene, and (iii) calculating an empirical gene-level P-value based on the proportion of times the observed test statistic is smaller than the simulated null statistics across all simulations (Liu *et al.* 2010).

PEGASUS: The main innovation in PEGASUS is using an analytical approach to compute gene-level P-values of observed gene scores according to a null distribution modeling LD (Figure 1D). Consider a gene (defined as the gene boundaries ±50 kb to include regulatory regions; the buffer of 50 kb can be varied) with n SNPs. Suppose the P-values for SNPs within the gene boundaries are $\{p_1, p_2, \ldots, p_n\}$ Let $x_i = F^{-1}(p_i)$, where F^{-1} is the inverse of the cumulative distribution function (CDF) $\chi^2_{\text{d.f.}=1}$. At the gene level, we are interested in the observed value q, defined as the sum of the correlated $\chi^2_{\text{d.f.}=1}$ variables within a gene (Equation 4):

$$q = \sum_{i=1}^{n} x_i. \tag{4}$$

Our model for q is as follows: Let $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ be an n-element multivariate normal vector with mean $\boldsymbol{\mu} = \mathbf{0}$ and positive definite covariance matrix $\sum$, where $\sum_{ij}$ is the LD between SNP i and SNP j and $\sum_{ii} = 1$. The quadratic form in the random variables $X_1, X_2, \ldots, X_n$ associated with an $n \times n$ symmetric matrix $A = (a_{ij})$ is defined as $Q(\mathbf{X}) = Q(X_1, X_2, \ldots, X_n) = \mathbf{X}'\mathbf{A}\mathbf{X} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} X_i X_j$ (Mathai and Provost 1992). The quadratic form $Q(\mathbf{X}) = \mathbf{X}'\mathbf{A}\mathbf{X}$ has the following representation (Equation 5), where $\boldsymbol{\lambda}$ are the eigenvalues of $\sum \mathbf{A}$ and $\mathbf{U}$ are mutually independent standard normal variables (Mathai and Provost 1992):

$$Q(\mathbf{X}) = \sum_{i=1}^{n} \lambda_i U_i^2. \tag{5}$$

$Q = \sum_{i=1}^{n} X_i^2$ follows the same distribution as Equation 5, and so the characteristic function of $Q(\mathbf{X})$ can be inverted to find the CDF of the null distribution accounting for empirical LD, which can be numerically integrated at the observed value (q) to find the gene-level P-value (p_g), $\text{Prob}(Q > q)$ (Mathai

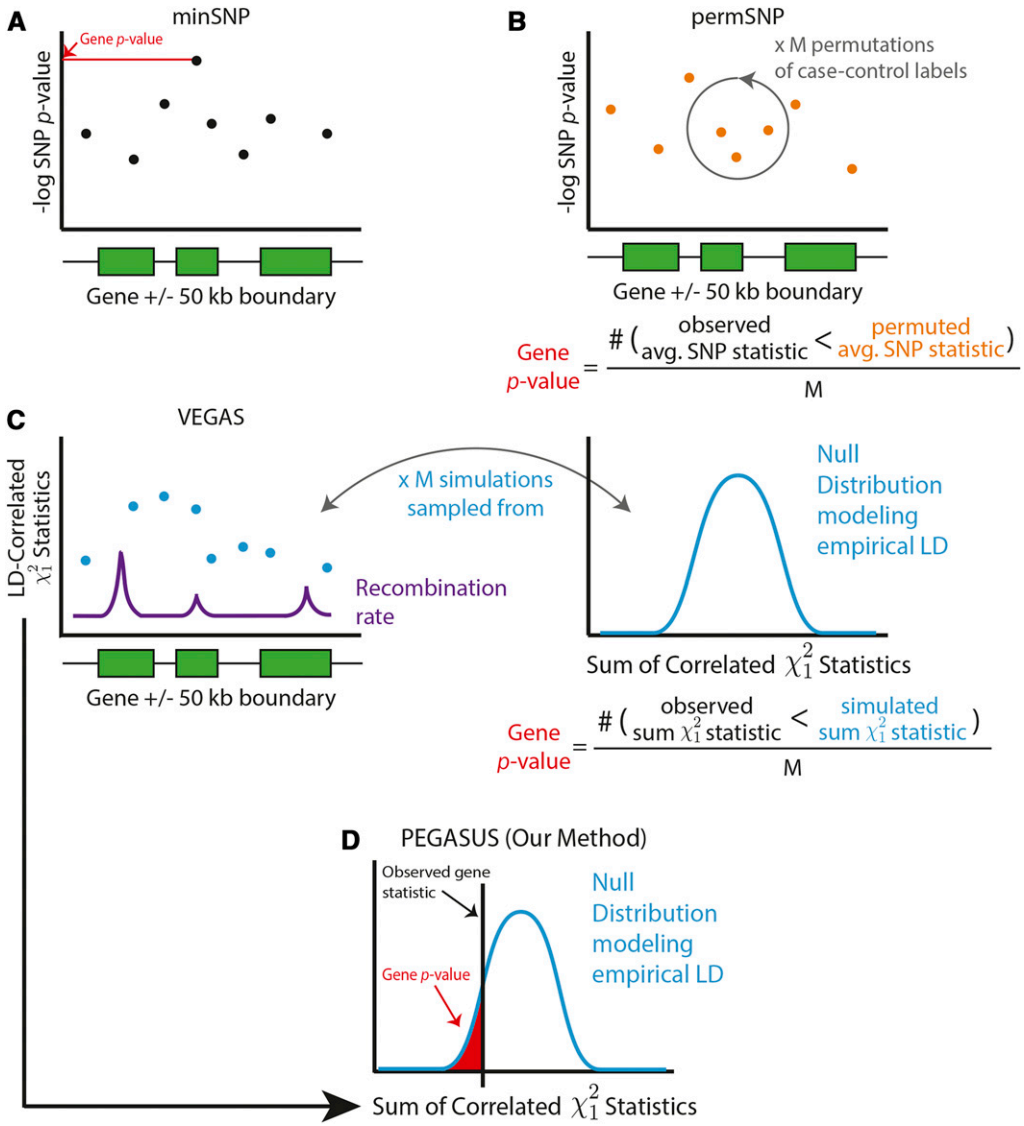


Figure 1 Schematic representations of PEGASUS and the three other methods—minSNP, permSNP, and VEGAS—assessed in this study. (A) minSNP defines the gene score to be the lowest of the SNP-level *P*-values within the gene observed in a GWA study. (B) permSNP (Ballard *et al.* 2010) performs *M* permutations of case–control labels in genotype data, recomputes GWA *P*-values for each SNP for each permuted data set, averages SNP *P*-values within each gene, and computes an empirical gene *P*-value based on the number of times the observed gene *P*-value is lower than permuted *P*-values. (C) VEGAS performs multivariate normal simulations from a null distribution of χ^2_1 statistics where the χ^2_1 statistics are correlated by empirical LD calculated from genotype data. *M* simulations are performed, the null statistics are summed within each gene and the empirical gene *P*-value is the number of times the observed χ^2_1 statistic is lower than the permuted χ^2_1 statistic. (D) In PEGASUS, for each gene, we numerically integrate the distribution of the sum of correlated χ^2_1 statistics at the observed gene statistic to determine the gene score. We also assess the performance of SKAT (Wu *et al.* 2010, 2011), which is not depicted here. SKAT uses a multiple linear/logistic regression framework, where genotypes for variants in a gene set and covariates are regressed onto phenotype to generate gene scores.

and Provost 1992). The numerical integration is implemented in the R package CompQuadForm (Duchesne and Lafaye De Micheaux 2010). The LD (covariance) matrix Σ is calculated using the --r flag (correlation) in PLINK (Purcell *et al.* 2007). In contrast, VEGAS (Liu *et al.* 2010) draws samples from the multivariate normal distribution with variance equal to the LD matrix, which are then summed to obtain an approximation of the *P*-value. Software to run PEGASUS is available at https://github.com/ramachandran-lab/PEGASUS. Empirical LD can be calculated using the 1000 Genomes Phase 3 data set (Auton *et al.* 2015) (release date: November 2014) as references. These data contain 2426 individuals in five superpopulations: East Asians, Europeans, Africans, South Asians, and admixed Americans.

Connection between SKAT and PEGASUS tests: As shown in Text S1, the SKAT and PEGASUS null distributions are mixtures of chi-square distributions. Mixture proportions for the SKAT null distribution are the eigenvalues of the matrix $\sigma^2[(\mathbf{I}-\mathbf{P})\mathbf{K}]$, where $\mathbf{P}=\mathbf{C}(\mathbf{C}^\mathbf{T}\mathbf{C})^{-1}\mathbf{C}^\mathbf{T}$ is a projection matrix dependent on the covariate matrix $\mathbf{C}$ and $\mathbf{K}=\mathbf{GWG}'$ is a kernel matrix dependent on the genotype matrix $\mathbf{G}$ and a diagonal matrix of weights $\mathbf{W}$. For the PEGASUS null distribution, mixture proportions are given by the eigenvalues of the LD matrix $\sum$. If no covariates are considered and the variant weights are uniform ($w_j=1$) for all variants, the SKAT null distribution becomes a mixture of chi-square distributions with mixture proportions given by the eigenvalues of the $\mathbf{K}=\mathbf{GWG}'$ matrix, which is a variance–covariance matrix similar to the PEGASUS LD matrix $\sum$. Thus, under these circumstances, the two tests give similar results. However, PEGASUS requires only summary statistics and is a better choice when genotype data are not available for analysis.

GWA study replication

To further assess the robustness of our method, PEGASUS, we attempted to replicate gene hits ($p_g < 2.8 \times 10^{-6}$ or 0.05 divided by approximately 18,000 genes tested) generated by PEGASUS for four data sets [bipolar disorder (BIP), Crohn's disease (CD), rheumatoid arthritis (RA), and type 2 diabetes (T2D)], using genotype data from the WTCCC (WTCCC 2007). For the replication study, we carried out PEGASUS on these four WTCCC data sets (our "replication cohort") and compared the top genes found in our "discovery" data sets (Franke *et al.* 2010; Stahl *et al.* 2010; Sklar *et al.* 2011; Morris *et al.* 2012) to those found in the WTCCC data sets. We note that this is not an independent replication study since the WTCCC data sets were included in the discovery cohorts; cases from the WTCCC data sets comprised at most 38% of the cases included in the discovery cohorts (Table S2). The replication data sets consist of ~2000 cases for each disease and ~3000 shared controls recruited from the United Kingdom and genotyped on the Affymetrix 500K GeneChip (WTCCC 2007).

Eight quality control steps were carried out for each of the four WTCCC data sets. Steps 1–7 were carried out using PLINK (Purcell *et al.* 2007) (version 1.07):

1. Markers with minor allele frequency <1% were removed.
2. Loci with a call rate ≤95% across individuals were removed.
3. Individuals with at least 5% missingness across all loci were removed.
4. Loci not in Hardy–Weinberg equilibrium were removed (P-value threshold of 10^{-5}).
5. Individuals were pruned based on inbreeding coefficient ($F \geq 0.05$ or $F \leq -0.025$).
6. Duplicate individuals were removed (one individual for each pair with identity by state ≥ 95%).
7. Related individuals were removed (one individual for each pair with $\hat{\pi} > 0.0175$).
8. Individuals determined to be outliers by principal component analysis were removed. SmartPCA from the EIGENSOFT (Price *et al.* 2006) software package (version 4.0.2) was used to do PCA with outlier removal. Five iterations of outlier removal were performed with the outlier σ threshold = 6.

We conducted GWA analysis using PLINK (Purcell *et al.* 2007) (version 1.07) on the WTCCC data sets. SNP-level P-values were determined by logistic regression of disease state onto minor allele dosage, using the top four principal components as covariates in the logistic regression to control for ancestry.

GWA study simulation

To compare how well minSNP, SKAT, VEGAS, and PEGASUS can recover causal genes, we conducted a GWA study for a simulated complex phenotype with known genetic architecture based on the approach outlined in Wojcik *et al.* (2015) and applied these four methods to the simulated data (Figure S13). To choose causal genes, we picked four pathways with >20 genes each at random from the KEGG pathway database (Kanehisa 1997; Kanehisa *et al.* 2012). For each pathway, we randomly sampled 20% of its genes, resulting in 54 total causal genes. We ran Tagger (Haploview, Version 4.3) (Barrett *et al.* 2005) on each gene to find independent tag SNPs ($r^2 < 0.2$), using the WTCCC controls ($N = 2900$ individuals) as reference individuals to calculate LD. For each of the 54 causal genes, we chose 1, 2, or 5 tag SNPs to be associated with the phenotype, giving 123 total causal SNPs. All the chosen SNPs in each gene were randomly assigned an effect size of either 1.2 or 2 to simulate a range of effect sizes.

Using software from Wojcik *et al.* (2015), we then calculated a per-individual liability score for each individual (WTCCC controls served as our simulated cases and controls) from a model of additive genetic effects by summing the effect size s of each SNP multiplied by the minor allele dosage X at the SNP over all n SNPs (Equation 6). A "wiggle" (ε) was added to each raw liability score (Equation 7) to allow the cases and controls to overlap in their liability score distributions:

$$\text{raw liability score} = \sum_{i=1}^{123 \text{ total causal SNPs}} s_i X_i \tag{6}$$

$$\begin{aligned} \text{wiggled score} &= \text{raw liability score} + \epsilon, \\ \text{where } \epsilon &\sim N(0.1, 10). \end{aligned} \tag{7}$$

Phenotype was assigned to each individual based on the mean of 100 deviates from the binomial distribution with probability of success equal to the probability of the wiggled score from the logistic distribution, which we obtained by applying the logistic function to the wiggled score.

We then conducted GWA analysis using PLINK (Purcell *et al.* 2007) (version 1.07) on the WTCCC controls and the simulated phenotypes. SNP-level P-values were determined by logistic regression of minor allele dosage onto disease state. We used the top four principal components, determined by applying smartPCA (Price *et al.* 2006) to the genotypes, as covariates in the logistic regression to control for ancestry. To simulate spurious associations between SNPs and our associated phenotype, we added 20% of significant SNP P-values (144 new SNPs total) from an existing GWA study on CD (Franke *et al.* 2010) to our simulated GWA P-values; these spuriously associated SNPs did not overlap with SNPs already associated with simulated phenotype. By "spuriously associated" SNPs, we mean SNPs that achieve genome-wide significance (P-value $< 5 \times 10^{-8}$) but are not discussed or selected for replication studies, eQTL analysis, or other downstream analyses due to filtering steps. Such SNPs may be excluded based on criteria such as failure to achieve significance within a majority of the individual cohorts analyzed in a meta-analysis (Anderson *et al.* 2011), location within regions with complex LD or complex association patterns with the trait such as the MHC or *TNFAIP3*

regions for RA (Stahl *et al.* 2010), or *P*-value thresholds based on additional *in silico* analyses such as GRAIL (Raychaudhuri *et al.* 2009; Franke *et al.* 2010). Since the true causal genes underlying the simulated phenotype are known, we are able to measure true positive rate (TPR) and false positive rate (FPR) for each gene score method and used the following gene score thresholds: $\{7.5\times10^{-1},\ 5\times10^{-1},\ 2.5\times10^{-1},\ 10^{-1},\ 7.5\times10^{-2},\ 5\times10^{-2},\ 2.5\times10^{-2},\ 10^{-2},\ \ldots,\ 7.5\times10^{-16}, 5\times10^{-16}, 2.5\times10^{-16}\}$ We find that our simulation results are robust to varying percentages of spuriously associated SNPs added and the GWA data set used (Figure S14).

Pathway analysis

We performed pathway analysis with HotNet2 (Leiserson *et al.* 2015), a topology-based method for finding significantly mutated subnetworks within protein–protein interaction networks, originally developed for analyzing somatic mutation data from cancer data sets. HotNet2 uses directed heat diffusion along interaction networks where every gene, represented by nodes in the network graph, has a "heat score" based on its gene score. We used negative log-transformed gene scores generated by PEGASUS, VEGAS, and minSNP as heat scores in HotNet2.

We use HotNet2 to find gene interaction subnetworks containing genes that we have highest confidence are truly associated with the phenotype. We found that HotNet2 does not perform well when too many genes are assigned similar heat scores, as will happen when the majority of genes have insignificant *P*-values. Thus, following the approach used in earlier applications of HotNet2 to somatic mutations in cancer, we assigned heat scores of zero to genes that we have low confidence are truly associated with the phenotype (Leiserson *et al.* 2015). To identify low-confidence genes, we compute a hard threshold based on local false discovery rates (lFDR) for the gene *P*-values. In a disease association setting, lFDR is the probability that a gene is not associated with the phenotype given its corresponding observed *P*-value; thus, $1-$ lFDR can be thought of as "confidence." When plotting confidence against gene scores (Figure S5 and Figure S6), an "elbow" or inflection point typically corresponds to a sharp drop in confidence; therefore, the inflection point is a natural choice for a gene score threshold. We compute the lFDRs for the gene scores using the twilight R package (Scheid and Spang 2005) (version 1.44.0). We calculate lFDR for minSNP, VEGAS, and PEGASUS gene scores and then determine a cutoff at the first elbow or inflection point in the graph of $1-$ lFDR against gene scores where possible (Figure S5 and Figure S6). If the graph has no elbow point, as in the minSNP lFDR curves, we used the gene score corresponding to an lFDR cutoff of 0.05 (Figure S4). Since the cutoffs for the minSNP scores were greater than those calculated using PEGASUS and VEGAS gene scores, we ran HotNet2 twice: once using the PEGASUS lFDR threshold and once with the higher minSNP lFDR threshold. We assessed significance of HotNet2 results for each run by permuting the heat scores on genes to find a *P*-value for the number of subnetworks containing $\geq k$ genes, as reported by HotNet2 (Leiserson *et al.* 2015).

HotNet2 analysis was performed using the HINT interaction network (Das and Yu 2012). Runs that had multiple *P*-values ≤ 0.05 of varying size k in the permutation test were further studied, for example, by annotation using the Genome-wide Repository of Associations between SNPs and Phenotypes (GRASP) GWA study catalog (Leslie *et al.* 2014) to determine significance of the genes in previous GWA studies, along with functional annotations and literature searches.

Data availability

GWA *P*-values analyzed here (Table 2) can be accessed at the URLs given in Table S3. WTCCC data (WTCCC 2007) used in power simulations and replication studies can be accessed through the Wellcome Trust Case Control Consortium: http://www.wtccc.org.uk/. PEGASUS software and LD reference data are available online at the following URL: https://github.com/ramachandran-lab/PEGASUS. HotNet2 (Leiserson *et al.* 2015) software is available online at the following URL: https://github.com/raphael-group/hotnet2.

Results

Performance comparison of PEGASUS against minSNP, permSNP, SKAT, and VEGAS

We compared PEGASUS against minSNP (Torkamani *et al.* 2008; Fehringer *et al.* 2012; Gelernter *et al.* 2015; Hu *et al.* 2015), permSNP (Wang *et al.* 2007; Eleftherohorinou *et al.* 2009; Ballard *et al.* 2010; Christoforou *et al.* 2014; Evangelou *et al.* 2014; Backes *et al.* 2016), SKAT (Wu *et al.* 2010, 2011), and VEGAS (Liu *et al.* 2010) (Figure 2 and Figure S15), using several metrics to evaluate the different scores.

We find that, for all 12 GWA data sets analyzed, minSNP gene scores are almost always smaller than PEGASUS gene scores (Figure 2A and Figure S1). We also find that minSNP gene scores show a clear dependence on gene length; as the number of SNPs in a gene increases, the minSNP gene score decreases for all data sets analyzed. In contrast, PEGASUS gene scores do not show this trend (Figure 2A, Figure S2, and Figure S16).

We tested whether two corrections to minSNP gene scores mitigated its bias with gene length: (i) calculating *P*-values for minSNP gene scores from the Beta(1, no. of SNPs in gene) distribution (note that minSNP can be thought of as the first-order statistic for SNP *P*-values) and (ii) multiplying minSNP gene scores by the number of SNPs in a gene. Both corrections resulted in gene scores that decrease with increasing gene length (Figure S7 and Figure S8).

Since permSNP requires genotype data and is computationally expensive, permSNP was carried out only on the ALL data set (Xu *et al.* 2013), using 10,000 permutations; thus, there is large variation in PEGASUS gene scores at a

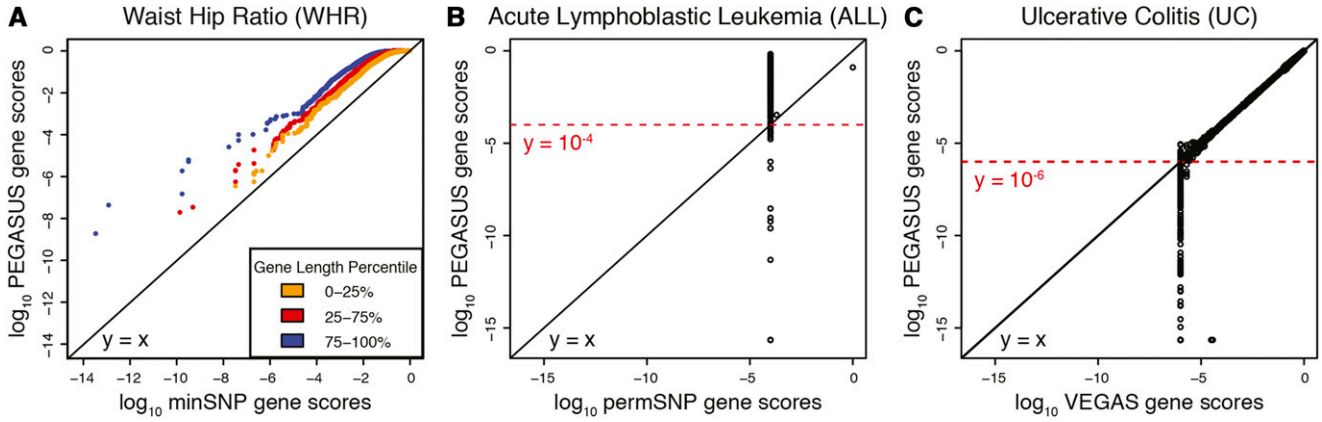


Figure 2 Quantile–quantile plots comparing gene scores produced by PEGASUS against those produced by minSNP, permSNP, and VEGAS. (A) Quantile–quantile plots of PEGASUS gene scores *vs.* minSNP gene scores. Each point represents a gene and is colored yellow, red, or blue based on gene length percentile, 0–25%, 25–75%, and 75–100%, respectively. The phenotype used is waist–hip ratio adjusted for body mass index (WHR). minSNP gene scores are smaller than PEGASUS gene scores and decrease with increasing number of SNPs in a gene. The deviations from $y = x$ show that minSNP scores are biased toward being smaller than PEGASUS scores, and this bias increases for increasing gene length (genes colored in blue and red). (B) Base-10 logarithm of PEGASUS gene scores *vs.* base-10 logarithm of permSNP gene scores for acute lymphoblastic leukemia (ALL). permSNP can determine gene scores only as low as the reciprocal of the number of permutations (10,000 in this case) whereas PEGASUS can determine gene scores as low as 2.22×10^{-16} (the numerical precision of R). Note that the minimum permSNP scores of 10^{-4} differ widely from their *P*-values computed by PEGASUS. (C) Base-10 logarithm of PEGASUS gene scores *vs.* base-10 logarithm VEGAS gene scores. Using 1 million simulations, the lowest gene scores output by VEGAS are 10^{-6}, while PEGASUS determines gene scores as low as 2.22×10^{-16}. In addition, for gene scores close to the reciprocal of the maximum number of simulations performed, VEGAS can return inaccurate gene scores compared to PEGASUS.

permSNP gene score of 10^{-4} (Figure 2B). Further, permSNP is extremely computationally costly: Carrying out permSNP on a random subset of 400 genes took ~6 hr. Thus, permSNP would be extremely computationally inefficient for analyzing a genome-wide human data set (~18,000–20,000 genes).

Since SKAT requires genotype data, SKAT was carried out on the ALL (Xu *et al.* 2013) and the WTCCC T2D (WTCCC 2007) data sets, using the top four principal components from PCA on these data sets as covariates. We find that PEGASUS gene scores and SKAT gene scores are correlated ($r = 0.44$ and $r = 0.49$; *P*-values $< 2 \times 10^{-16}$ and $< 2 \times 10^{-16}$ for ALL and T2D, respectively) for both data sets (Figure S15, A and B). We also find that PEGASUS gene scores and unweighted SKAT gene scores are correlated ($r = 0.96$ and $r = 0.94$; *P*-values $< 2 \times 10^{-16}$ and $< 2 \times 10^{-16}$ for ALL and T2D, respectively) (Figure S15, C and D).

Compared to VEGAS, our method has increased numerical precision when calculating gene scores (Figure 2C). Due to its underlying Monte Carlo simulations (10^6 by default), VEGAS does not calculate gene scores less than the reciprocal of the number of simulations. However, PEGASUS can evaluate gene scores to the machine precision of R, which is $\sim 10^{-16}$. In addition, VEGAS gene scores become inaccurate close to 10^{-6} due to the random nature of Monte Carlo simulations (Figure 2C), whereas PEGASUS does not have a stochastic element. We find that VEGAS produces less numerically precise gene scores than PEGASUS in all 12 data sets analyzed (Figure S3). We also find that PEGASUS runs twice as fast as VEGAS when using HapMap data (Frazer *et al.* 2007) (Phase 2) as references for LD.

Enrichment for known associations in real data

To assess how well minSNP and PEGASUS recover known GWA associations, we calculated the percentage of genes with significant minSNP and PEGASUS gene scores ($p_g < 2.8 \times 10^{-6}$) for the 12 phenotypes in this analysis that have been found to be significantly associated (SNP-level *P*-value $< 5 \times 10^{-8}$) with the disease or trait in GWA studies conducted with different genotype data (Figure 3). Taking known associations to be "true positives," we calculated positive predictive values (PPV) for every gene score for every disease. We find that in 10 of 12 data sets, significantly associated PEGASUS gene hits ($p_g < 2.8 \times 10^{-6}$) have higher PPV than minSNP gene hits by as much as 2.8-fold, showing that PEGASUS gene hits are enriched for known associations in comparison to minSNP (Figure 3). For the remaining two disorders, attention-deficit/hyperactivity disorder (ADHD) and major depressive disorder (MDD), minSNP identifies significantly associated genes ($p_g < 2.8 \times 10^{-6}$) that have not been found in other GWA studies while PEGASUS does not report any findings (Figure 3).

Replication of PEGASUS gene hits in WTCCC data

We attempted to replicate gene hits ($p_g < 2.8 \times 10^{-6}$) generated by PEGASUS for four data sets (BIP, CD, RA, and T2D) for which we have genotype data from the WTCCC (WTCCC 2007). We note that our replication cohorts (WTCCC) were included in the discovery cohorts, and thus this is not an independent replication. We find that we are able to replicate up to 57.2% of gene hits in the case of RA and as low as 0 gene hits in the BIP data set (Table S2). We note that the four meta-analyses we consider our "discovery cohorts" are

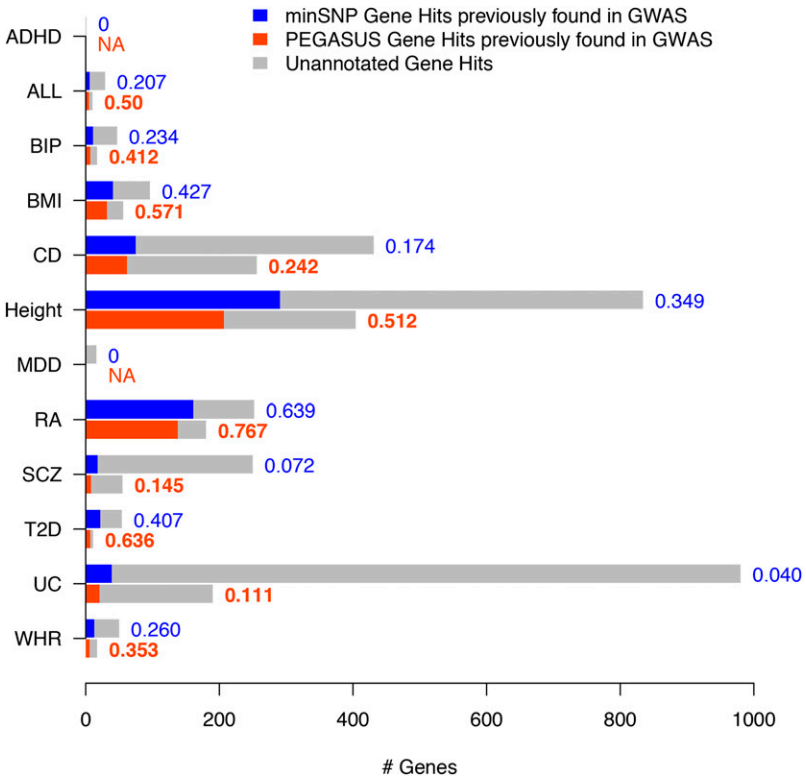


Figure 3 PEGASUS gene hits are enriched for known GWA study associations compared to minSNP gene hits. Shown are the numbers of minSNP gene hits (blue) and PEGASUS gene hits (orange) that contain known GWA study associations and gene hits not previously found in GWA studies (gray) for 12 GWA study data sets. To the right of each bar are positive predictive values (PPV) for each gene score for every data set; where possible, boldface type indicates the gene score with the highest PPV for each disease and "NA" means that PPV is undefined, which occurs when there are zero gene hits. A gene hit is a gene with a score of $< 2.8 \times 10^{-6}$, and known GWA study associations are genes containing genome-wide significant SNPs in GWA studies conducted with different data sets from the 12 data sets analyzed here. We find that PEGASUS gene hits have as much as 2.8-fold higher PPV than minSNP gene hits.

composed of much larger sample sizes than our replication data sets and thus had greater power to identify associated variants. Further, we do not necessarily follow the same steps for quality control and ancestry correction in our GWA study as did the meta-analyses reanalyzed here (see *Materials and Methods*), which may explain our low percentage of replicated significant genes for BIP and CD.

Power analysis using simulated GWA data

To compare how well minSNP, SKAT, VEGAS, and PEGASUS can recover causal genes, we conducted gene-level tests for association, using these methods on a GWA study for a simulated phenotype calculated from a model of additive genetic effects. PEGASUS and VEGAS outperform minSNP with a 30% TPR when the FPR is fixed at 1%, and PEGASUS and VEGAS outperform SKAT with 28% higher TPR when FPR is fixed at 1% (Figure 4). We find that minSNP, when applied to the GWA data sets in Table 2, outputs high numbers of significant genes (genes with $p_g < 2.8 \times 10^{-6}$). For example, as many as 5.5% of all genes (lower bound: 0.01%) are below the Bonferroni-corrected threshold for significance when using minSNP across the 12 data sets we analyzed (Table S1). In our simulation, we find that the high FPR in minSNP is due to genes with spurious SNP association *P*-values added in as part of the simulation. This suggests that some minSNP gene hits in observed GWA studies are spurious associations as well. By correcting for LD, PEGASUS and VEGAS are able to ignore these false positives (Figure 4). At very low FPR ($< 0.34\%$), SKAT is the most sensitive method, but SKAT has lower sensitivity overall. This could be a desirable feature if one is looking for a small number of reliable gene hits. However, for pathway/network analysis, it is useful to obtain high sensitivity, as the pathway/network information will help reduce the remaining false positives. We were unable to add in spurious SNP association *P*-values for the SKAT method since this test requires genotype data; however, inclusion of these SNPs could only decrease the performance of SKAT in simulation.

The downstream effect of gene scores on pathway analysis

Using gene scores generated from minSNP, VEGAS, and PEGASUS as input to HotNet2 (Leiserson *et al.* 2015), we performed pathway analysis on each of the 12 GWA data sets. We find significantly associated gene subnetworks for ADHD, ulcerative colitis (UC), and waist–hip ratio adjusted for body mass index (WHR). Figure 5 shows a selection of subnetworks containing known or biologically plausible gene associations for these three phenotypes based on previous GWA or functional studies. Other significantly associated subnetworks can be found in Figure S9 and Figure S12.

PEGASUS identifies multiple subnetworks containing genes with known associations to each phenotype of interest. Some of these subnetworks are not found when using minSNP and VEGAS gene scores as input to HotNet2 (Figure S10 and Figure S11). We also identify subnetworks associated with ADHD (Figure 5, D–F), a disorder for which GWA studies have not identified any SNPs with genome-wide significance. We detail our findings for each disease or trait below.

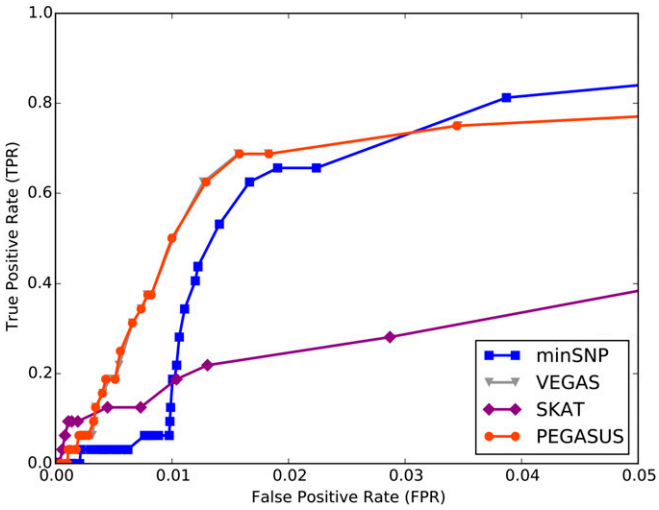


Figure 4 Receiver operating characteristic (ROC) curves from GWA conducted with simulated phenotypes show gene scores controlling for LD achieve higher true positive rates at low false positive rates. We performed a GWA study for a simulated phenotype with known underlying true causal genes (see *Materials and Methods*) and determined true positive rate (TPR; genes truly associated with phenotype that were identified as such) and false positive rate (FPR; genes identified as causal by a gene score method that were not truly associated with the simulated phenotype) for minSNP, VEGAS, SKAT, and PEGASUS for various gene score thresholds (see *Materials and Methods*). We find that PEGASUS and VEGAS, which control for pairwise correlations between SNPs within genes, outperform minSNP and SKAT with higher TPRs at very low FPRs.

UC: Inflammatory bowel disease (IBD), an inflammatory disease of the gastrointestinal tract, has two major subtypes: UC and CD. IBD is hypothesized to result from dysregulated T-cell immune responses to commensal enteric bacteria in the gut that develop in individuals who are genetically predisposed to the disease. Environmental factors also play an important role in triggering onset or recurrence of symptoms (Sartor 2006; Lee *et al.* 2012). UC is characterized by superficial, ulcerating inflammation that is limited to the colon (Christophi *et al.* 2012).

HotNet2 analysis using PEGASUS scores identifies a subnetwork containing several genes in JAK2-STAT signaling pathways as associated with UC disease state (Figure 5A). The subnetwork in Figure 5A contains the genes *JAK2*, *IL12RB2*, *IFNG*, *PTPN2*, and *STAT4*, which all have genome-wide significant SNP hits (SNP *P*-value $< 5 \times 10^{-8}$) in GWA studies for IBD (Duerr *et al.* 2006; Jostins *et al.* 2012). The gene *IFNG* also has genome-wide significant SNPs in GWA studies for ulcerative colitis conducted with different data sets from the data set used in this study (2.5×10^{-12} and 4.2×10^{-12}) (Silverberg *et al.* 2009; McGovern *et al.* 2010). The following genes shown in this subnetwork have also been significantly associated (SNP *P*-value $< 5 \times 10^{-8}$) with CD (CD alone or in concert with psoriasis or celiac disease) in other GWA studies: *JAK2*, *IL12RB2*, *SOCS1*, and *PTPN2* (Raelson *et al.* 2007; WTCCC 2007; Barrett *et al.* 2008; Franke *et al.* 2010; Festen *et al.* 2011; Ellinghaus *et al.* 2012). Two closely related pro-inflammatory cytokine signaling pathways involve many genes shown in this subnetwork: the interleukin (IL)-23/type 17 helper T-cell (T_H17) signaling pathway and the IL-12/type 1 helper T-cell (T_H1) signaling pathway. Both signaling pathways ultimately result in cytokine-mediated gut destruction (Wang *et al.* 2010; Parkes *et al.* 2013).

Similar pathways centered around IL-12, IL-23, and JAK2-STAT signaling have been manually compiled based on GWA studies for CD and IBD (Wang *et al.* 2010; Parkes *et al.* 2013). In addition, the subnetwork (Figure 5A) contains the gene *PTPN2*, which encodes protein tyrosine phosphatase nonreceptor type 2 (PTPN2) and has been shown to regulate autophagy in human intestinal epithelial cells; knockdown of *PTPN2* caused impaired autophagosome formation and dysfunctional autophagy that eventually resulted in increased apoptosis of intestinal cells in response to IFNG and tumor necrosis factor-α (Scharl *et al.* 2012). This subnetwork illustrates interactions between multiple genes involved in the immune response to pathogens that may underlie the pathology of ulcerative colitis.

The subnetwork we identify in Figure 5B shows interactions between the human leukocyte antigen (HLA) class I genes and transporter associated with antigen processing (TAP) genes. These genes are thought to underlie IBD pathology as well as other immune-mediated disorders such as psoriasis and ankylosing spondylitis (Parkes *et al.* 2013). Finally, HotNet2 reports a significantly associated subnetwork containing genes that are part of the tumor necrosis factor (TNF) signaling pathway (Figure 5C). TNF signaling results in activation of nuclear factor kappa-light-chain enhancer of activated B cells (NF-κB), which is a known inflammatory response in IBD (Anderson *et al.* 2011). Additional significant gene subnetworks are shown in Figure S12.

We emphasize that neither the JAK2-STAT subnetwork (Figure 5A) nor the HLA class 1 and TAP genes (Figure 5B) are found when using VEGAS gene scores as input to HotNet2 (Figure S11), demonstrating the importance of high-precision gene scores in downstream analysis.

ADHD: ADHD is a highly heritable neuropsychiatric disorder characterized by the following traits: inattention, hyperactivity, and impulsivity (Franke *et al.* 2009). It is thought to be a very complex multifactorial trait and, despite high heritability estimates (76%) (Neale *et al.* 2010), it has been difficult to find genes underlying the phenotype. GWA studies with very large sample sizes have been performed, yet no variants have reached genome-wide significance (Franke *et al.* 2009).

Using PEGASUS gene scores as input to HotNet2, we find five significantly associated gene subnetworks associated with ADHD (Figure 5 and Figure S9). Figure 5D includes interactions between multiple genes previously associated with cognition-related traits and neurodevelopmental disorders. *RPS6KA5*, *PDZD2*, and *RAI1* are associated with years of education (GWA SNP *P*-values of 9.2×10^{-4}, 5.9×10^{-4}, and 3.6×10^{-5}, respectively) (Rietveld *et al.* 2013). The genes *RPS6KA5*, *PDZD2*, and *ST3GAL3* have also been associated

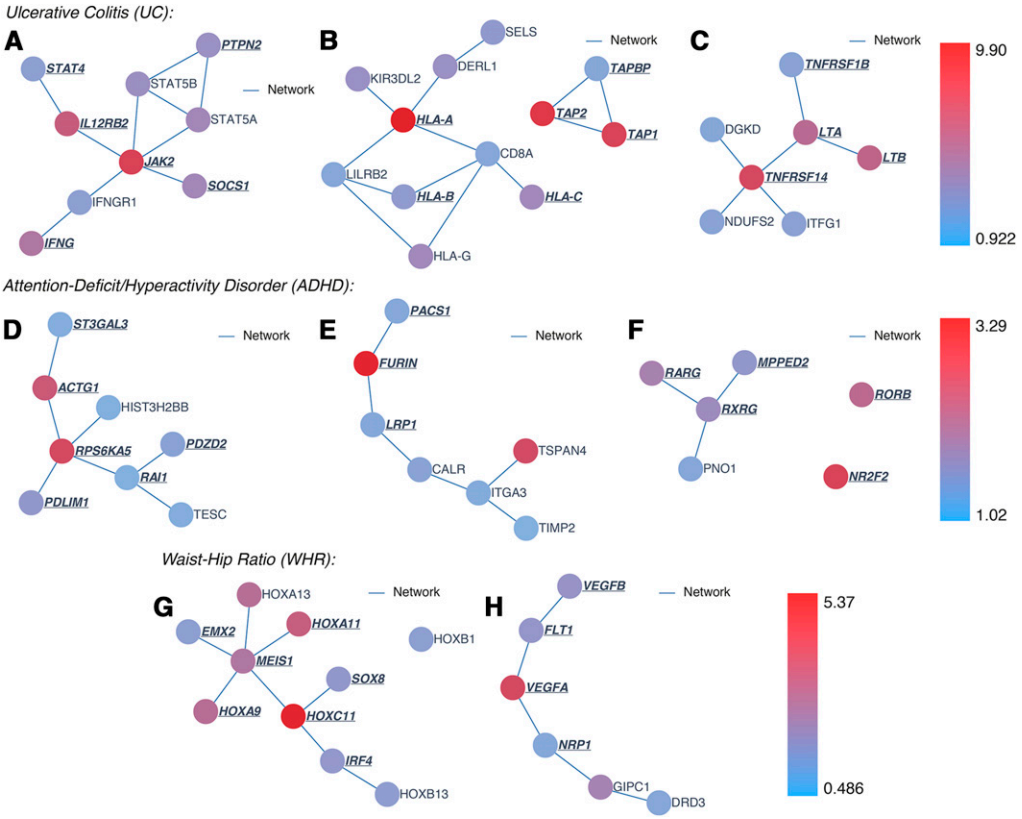


Figure 5 Subnetworks for ulcerative colitis (A–C), attention-deficit/hyperactivity disorder (D–F), and waist–hip ratio adjusted for body mass index (G and H) from significant runs of HotNet2 (Leiserson *et al.* 2015) ($p \leq 0.05$ for multiple subnetwork sizes), using PEGASUS gene scores as input. Circles represent genes in each subnetwork and are colored by heat score (negative log-transformed PEGASUS gene scores); the color bar indicates the lowest heat score (blue or "cold" genes) and the highest heat score (red or "hot" genes) in each subnetwork for a given phenotype. Lines between genes indicate a direct gene–gene interaction from the HINT database (Das and Yu 2012). Gene names that are underlined, in boldface type, and italicized represent genes that have been previously associated with the ulcerative colitis (Duerr *et al.* 2006; Raelson *et al.* 2007; WTCCC 2007; Barrett *et al.* 2008; Silverberg *et al.* 2009; Franke *et al.* 2010; McGovern *et al.* 2010; Anderson *et al.* 2011; Festen *et al.* 2011; Ellinghaus *et al.* 2012; Jostins *et al.* 2012; Scharl *et al.* 2012; Parkes *et al.* 2013), attention-deficit/hyperactivity disorder (Wan *et al.* 1998; Maden 2007; Davis *et al.* 2008; Naka *et al.* 2008; McGrath *et al.* 2009; Need *et al.* 2009; Chen *et al.* 2010; Cirulli *et al.* 2010, 2012; Fuentealba *et al.* 2010; Neale *et al.* 2010; Hu *et al.* 2011; Luciano *et al.* 2011; Tang *et al.* 2011; De Jager *et al.* 2012; Rivière *et al.* 2012; Schuurs-Hoeijmakers *et al.* 2012; Peixoto and Abel 2013; Rietveld *et al.* 2013), and waist–hip ratio (Cantile *et al.* 2003; Eguchi *et al.* 2008, 2011; Guth *et al.* 2009; Hagberg *et al.* 2010; Heid *et al.* 2010; Siervo *et al.* 2012; Tchkonia *et al.* 2013; Karpe and Pinnick 2015) phenotypes in GWA or functional studies.

($3.4 \times 10^{-5} \leq p \leq 2 \times 10^{-4}$) with performance on multiple tests of cognitive function and memory in previous GWA studies (Need *et al.* 2009; Cirulli *et al.* 2010, 2012; Luciano *et al.* 2011; De Jager *et al.* 2012). In particular, *RPS6KA5*, which encodes ribosomal protein S6 kinase alpha-5, is part of a pathway involved in brain-derived neurotrophic factor (BDNF)/neurotrophin signaling; BDNF and other neurotrophins are important in neural development, learning, and memory and are implicated in neurodegenerative diseases such as Huntington's, Alzheimer's and Parkinson's (Tang *et al.* 2011). *RPS6KA5* is also thought to be a part of the upstream pathway involved in learning-dependent chromatin remodeling, which is important in long-term memory formation (Peixoto and Abel 2013). In addition, the gene *PDLIM1* has been found to have strong maternal transmission in trios where the child is affected with ADHD, and *PDLIM1* has been shown to play a role in Alzheimer's disease (Wang *et al.* 2012). A linkage study for intellectual disability found mutations in the *ST3GAL3* gene (Hu *et al.* 2011), and mutations in the *ACTG1* gene cause Baraitser–Winter syndrome, a developmental disorder affecting the face and brain (Rivière *et al.* 2012). Taken together, multiple studies suggest that genes in this subnetwork we identified using PEGASUS and HotNet2 are involved in regulatory processes that affect neural development, learning, and memory.

The second subnetwork (Figure 5E) contains *FURIN* and other genes that play important roles in cell trafficking processes that may be involved in the pathology of neuropsychiatric and cognitive disorders such as Alzheimer's disease and intellectual disability (Wan *et al.* 1998; Fuentealba *et al.* 2010; Schuurs-Hoeijmakers *et al.* 2012; Carlino *et al.* 2013). A third subnetwork (Figure 5F) contains genes that are likely involved in brain development (Maden 2007). Two additional subnetworks found by HotNet2 using PEGASUS scores contain genes that play regulatory roles in neurogenesis and responses to stress (Figure S9A) and genes associated with other social and behavioral abnormalities (Figure S9B). Additional information about genes in these subnetworks can be found in Text S2.

WHR adjusted for body mass index: WHR adjusted for body mass index (BMI) is a quantitative trait that measures body fat distribution. Both WHR and BMI are heritable traits (25–70% heritability), but mechanisms underlying body fat distribution are still unclear (Baker *et al.* 2005). WHR is a useful trait for predicting risk for T2D and heart disease since it accounts for waist and hip size, which both have associations with these traits (Heid *et al.* 2010). Increasing waist size is associated with increased risk for T2D and heart disease, but gluteal fat deposits play a protective role against T2D,

hypertension, and dyslipidemia (Heid *et al.* 2010; Shungin *et al.* 2015).

HotNet2 analysis with PEGASUS gene scores identifies two subnetworks with interactions between genes known to be associated with WHR that may shed light on the genetic determination of body fat distribution. Figure 5G displays a subnetwork of homeobox (HOX) genes, a family of transcription factors that play an important role in morphogenesis in animals (Zhang *et al.* 2007). The complete HOX gene network was found to be active in human white adipose tissue and fetal brown adipose tissue (Cantile *et al.* 2003). HOX genes contained in this subnetwork include *HOXA9*, *HOXC11*, *HOXA11*, and *EMX2*. Multiple studies of gene expression in human subcutaneous abdominal adipose tissue and gluteal adipose tissue found that the *HOXA9* gene has increased expression in abdominal adipose tissue; *MEIS1*, which encodes a HOX cofactor and is also contained in this subnetwork, was also expressed more in abdominal adipose tissue than in gluteal depots in men only (Karpe and Pinnick 2015). In contrast, *HOXC11* and *HOXA11* have increased gene expression in gluteal adipose tissue than in abdominal adipose tissue (Karpe and Pinnick 2015). *HOXA9* and *EMX2*, another homeobox gene, were induced after extreme weight loss following bariatric surgery (Tchkonia *et al.* 2013). The subnetwork shows *SOX8*, which encodes a transcription factor thought to play a role in development; mice deficient in *SOX8* develop surprisingly normally, but undergo a severe degeneration of adipose tissue as adult mice (Guth *et al.* 2009). Guth *et al.* (2009) posit that *SOX8* plays a role in adipocyte development, especially during replenishment of the adipocyte pool in adult mice. *IRF4* is also contained in this subnetwork and encodes interferon regulatory factor 4 (IRF4), which is part of a family of transcription factors that are involved in various immune functions including regulation of innate immunity via the Toll-like receptor (TLR) signaling pathway (Eguchi *et al.* 2011). Many IRF proteins including IRF4 are also expressed in preadipocytes and mature adipocytes; their function is to repress adipogenesis (Eguchi *et al.* 2008). Eguchi *et al.* (2011) find that knockout mice lacking *IRF4* in adipocytes display excess adiposity and resistance to lipolysis induced by catecholamines, fasting, or cold exposure, suggesting that *IRF4* plays an important role in transcriptional regulation of lipid handling in fat. These findings indicate that this subnetwork may play an important role in development of adipocytes and fat distribution.

The second subnetwork found by HotNet2 (Figure 5H) elucidates the interactions of *VEGFA*, which contains a known GWA study association for WHR (SNP *P*-value 1.38×10^{-10}) and *VEGFB*, which is moderately associated with WHR (SNP *P*-value 7.2×10^{-3}) (Heid *et al.* 2010). Additional information about this subnetwork can be found in Text S3.

We also conducted HotNet2 analysis using minSNP gene scores for all phenotypes studied here. Since minSNP gene scores are misleadingly small (Figure 2A, Figure S2, and Figure S16), single highly significant genes pull in many unrelated genes with low gene scores to create artifactual "star-shaped" subnetworks that are likely false positives (see Figure S10 for HotNet2 results using minSNP gene scores for ulcerative colitis).

Discussion

Here we present a new approach for identifying geno–phenotype associations from case–control data that combines a novel gene score, PEGASUS, with network-based analyses using HotNet2 (Leiserson *et al.* 2015). PEGASUS computes gene scores that measure the statistical association of a gene with a phenotype of interest and has multiple advantages over commonly used methods for generating gene scores.

First, PEGASUS analytically models LD among SNPs within genes, producing a computationally efficient method that yields precise results—gene scores as small as 2.22×10^{-16} (the machine precision of R). By modeling linkage disequilibrium, PEGASUS, like VEGAS (Liu *et al.* 2010), is sensitive to genes with multiple SNPs that are moderately associated with the phenotype of interest. Unlike existing methods, our gene scores are not biased by gene length or poor precision. In the future, our approach can be extended to combine SNP-level *P*-values within linkage blocks in contrast to the gene boundaries used in this study.

Second, we apply PEGASUS to 12 genome-wide association studies for complex diseases and traits and, in 10 of 12 studies, our significant gene scores ($p_g < 2.8 \times 10^{-6}$) enrich for genes known to be associated with the phenotypes of interest (Figure 3). In simulation studies, we find that modeling fine-scale LD (or pairwise correlations between SNPs within genes), as in the PEGASUS and VEGAS methods, produces gene scores with much higher true positive rates of identifying genes associated with the simulated phenotype than does using minSNP (Figure 4). Thus, PEGASUS can be applied after conducting a GWA study to prioritize genes for functional validation.

Third, because our approach precisely assesses the statistical association of a gene with a phenotype of interest, PEGASUS offers the opportunity to identify novel sets of interacting genes underlying complex phenotypes via pathway or network analysis with our gene scores as input. Complex phenotypes may be generated via mutations in a subset of genes in a predefined gene set [such as the gene sets used in gene set enrichment analysis (Subramanian *et al.* 2005)] or via crosstalk between gene sets or pathways. To identify novel subnetworks of genes associated with each of the complex phenotypes analyzed in this study, we used gene scores generated by PEGASUS as input to HotNet2 (Leiserson *et al.* 2015); this is the first application of HotNet2 to common genetic variation. In our network analyses, minSNP and VEGAS miss key subnetworks containing genes known to be associated with the phenotype of interest (Figure 5, A and B, ulcerative colitis).

Fourth, in ADHD, a disease for which large GWA studies ($n > 9543$ individuals) have not identified genome-wide significant associations (Franke *et al.* 2009; Neale *et al.* 2010),

we identify significant subnetworks of interacting genes ($p \leq 0.05$) that are involved in neural development and learning and cognition. As knowledge of the human interactome continues to improve, *post hoc* analysis of GWA studies using gene scores from PEGASUS in conjunction with HotNet2 has the potential to generate promising hypotheses for functional validation.

Finally, we argue that human genetics would benefit greatly if more GWA studies released *all* SNP-level *P*-values generated, instead of reporting a subset of *P*-values the authors consider to be of interest. The present situation, where only the genotypes are deposited in public repositories under managed access, makes it impossible to replicate the various filtering, quality control, ancestry correction, and other steps that lead from raw genotype calls to the handful of genome-wide significant SNPs reported in publications. Of course, participant confidentiality requirements may limit the public distribution of SNP *P*-values (Masca *et al.* 2011), but *P*-values could be released with the genotype data under a managed access model. If these summary statistics were routinely released, gene score methods like PEGASUS, network/pathway analyses like HotNet2 (Leiserson *et al.* 2015), and other computational innovations from the community could be more widely applied to yield new insight into the genomic underpinnings of complex diseases and traits.

Acknowledgments

We thank the Psychiatric Genomics Consortium, the Genetic Investigation of Anthropometric Traits Consortium, the International Inflammatory Bowel Disease Genetics Consortium, the Diabetes Genetics Replication and Meta-analysis Consortium, and the Broad Institute for making full genome-wide association (GWA) *P*-values data sets available for public download. We also thank Heng Xu, Virginia Perez-Andreu, and Jun J. Yang from the St. Jude Children's Research Hospital for providing full GWA *P*-values and genotype data from their multiethnic acute lymphoblastic leukemia GWA study (Xu *et al.* 2013) and for help with curating the raw genotype data. We gratefully acknowledge Max Leiserson and Jonathan Eldridge for assistance with HotNet2 analysis; Matt Reyna, Julia Palacios, and Lauren A. Sugden for helpful discussions; and Genevieve Wojcik for providing software and help with GWA simulations. We also thank Chris Cotsapas for helpful discussions. B.J.R. is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an Alfred P. Sloan Research Fellowship, U.S. National Science Foundation (NSF) grant IIS-1016648, an NSF CAREER award (CCF-1053753), and U.S. National Institutes of Health (NIH) grants R01HG007069 and R01CA180776. P.N. is supported by an Oliver Cromwell Gorton Arnold predoctoral fellowship from Brown University and by NSF CAREER award DBI-1452622 (to S.R.). S.R. is also supported by NIH grant R01GM118652, the Pew Charitable Trusts as a Pew Scholar in the Biomedical Sciences, and an Alfred P. Sloan Research Fellowship.

Literature Cited

Anderson, C. A., G. Boucher, C. W. Lees, A. Franke, M. D'Amato *et al.*, 2011 Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat. Genet. 43: 246–252.

Auton, A., G. R. Abecasis, D. M. Altshuler, R. M. Durbin, D. R. Bentley *et al.*, 2015 A global reference for human genetic variation. Nature 526: 68–74.

Backes, C., B. Meder, A. Lai, M. Stoll, F. Rühle *et al.*, 2016 Pathway-based variant enrichment analysis on the example of dilated cardiomyopathy. Hum. Genet. 135: 31–40.

Baker, M., N. Gaukrodger, B. M. Mayosi, H. Imrie, M. Farrall *et al.*, 2005 Association between common polymorphisms of the proopiomelanocortin gene and body fat distribution: a family study. Diabetes 54: 2492–2496.

Ballard, D. H., J. Cho, and H. Zhao, 2010 Comparisons of multimarker association methods to detect association between a candidate region and disease. Genet. Epidemiol. 34: 201–212.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly, 2005 Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.

Barrett, J. C., S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr *et al.*, 2008 Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. 40: 955–962.

Buch, S., F. Stickel, E. Trépo, M. Way, A. Herrmann *et al.*, 2015 A genome-wide association study confirms PNPLA3 and identifies TM6SF2 and MBOAT7 as risk loci for alcohol-related cirrhosis. Nat. Genet. 47: 1443–1448.

Cantile, M., A. Procino, M. D'Armiento, L. Cindolo, and C. Cillo, 2003 HOX gene network is involved in the transcriptional regulation of in vivo human adipogenesis. J. Cell. Physiol. 194: 225–236.

Carlino, D., M. De Vanna, and E. Tongiorgi, 2013 Is altered BDNF biosynthesis a general feature in patients with cognitive dysfunctions? Neuroscientist 19: 345–353.

Chen, C.-M., H.-Y. Wang, L.-R. You, R.-L. Shang, and F.-C. Liu, 2010 Expression analysis of an evolutionarily conserved metallophosphodiesterase gene, Mpped1, in the normal and beta-catenin-deficient malformed dorsal telencephalon. Dev. Dyn. 239: 1797–1806.

Christoforou, A., T. Espeseth, G. Davies, C. P. D. Fernandes, S. Giddaluru *et al.*, 2014 GWAS-based pathway analysis differentiates between fluid and crystallized intelligence. Genes Brain Behav. 13: 663–674.

Christophi, G. P., R. Rong, P. G. Holtzapple, P. T. Massa, and S. K. Landas, 2012 Immune markers and differential signaling networks in ulcerative colitis and Crohn's disease. Inflamm. Bowel Dis. 18: 2342–2356.

Cirulli, E. T., D. Kasperavicite, D. K. Attix, A. C. Need, D. Ge *et al.*, 2010 Common genetic variation and performance on standardized cognitive tests. Eur. J. Hum. Genet. 18: 815–820.

Cirulli, E. T., T. J. Urban, S. E. Marino, K. N. Linney, A. K. Birnbaum *et al.*, 2012 Genetic and environmental correlates of topiramate-induced cognitive impairment. Epilepsia 53: e5–e8.

Daly, A. K., 2010 Genome-wide association studies in pharmacogenomics. Nat. Rev. Genet. 11: 241–246.

Das, J., and H. Yu, 2012 HINT: high-quality protein interactomes and their applications in understanding human disease. BMC Syst. Biol. 6: 92.

Davis, L. K., K. J. Meyer, D. S. Rudd, A. L. Librant, E. A. Epping *et al.*, 2008 Pax6 3′ deletion results in aniridia, autism and mental retardation. Hum. Genet. 123: 371–378.

De Jager, P. L., J. M. Shulman, L. B. Chibnik, B. T. Keenan, T. Raj *et al.*, 2012 A genome-wide scan for common variants affecting the rate of age-related cognitive decline. Neurobiol. Aging 33: 1017.e1–1017.e15.

Duchesne, P., and P. Lafaye De Micheaux, 2010 Computing the distribution of quadratic forms: further comparisons between the Liu-Tang-Zhang approximation and exact methods. Comput. Stat. Data Anal. 54: 858–862.

Duerr, R. H., K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg *et al.*, 2006 A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 314: 1461–1463.

Eguchi, J., Q.-W. Yan, D. E. Schones, M. Kamal, C.-H. Hsu *et al.*, 2008 Interferon regulatory factors are transcriptional regulators of adipogenesis. Cell Metab. 7: 86–94.

Eguchi, J., X. Wang, S. Yu, E. E. Kershaw, P. C. Chiu *et al.*, 2011 Transcriptional control of adipose lipid handling by IRF4. Cell Metab. 13: 249–259.

Eleftherohorinou, H., V. Wright, C. Hoggart, A.-L. Hartikainen, M.-R. Jarvelin *et al.*, 2009 Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. PLoS One 4: e8068.

Ellinghaus, D., E. Ellinghaus, R. P. Nair, P. E. Stuart, T. Esko *et al.*, 2012 Combined analysis of genome-wide association studies for Crohn disease and psoriasis identifies seven shared susceptibility loci. Am. J. Hum. Genet. 90: 636–647.

Evangelou, E., and J. P. A. Ioannidis, 2013 Meta-analysis methods for genome-wide association studies and beyond. Nat. Rev. Genet. 14: 379–389.

Evangelou, M., D. J. Smyth, M. D. Fortune, O. S. Burren, N. M. Walker *et al.*, 2014 A method for gene-based pathway analysis using genomewide association study summary statistics reveals nine new type 1 diabetes associations. Genet. Epidemiol. 38: 661–670.

Fehringer, G., G. Liu, L. Briollais, P. Brennan, C. I. Amos *et al.*, 2012 Comparison of pathway analysis approaches using lung cancer GWAS data sets. PLoS One 7: e31816.

Festen, E. A. M., P. Goyette, T. Green, G. Boucher, C. Beauchamp *et al.*, 2011 A meta-analysis of genome-wide association scans identifies IL18RAP, PTPN2, TAGAP, and PUS10 as shared risk loci for Crohn's disease and celiac disease. PLoS Genet. 7: e1001283.

Finucane, H. K., B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef *et al.*, 2015 Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 47: 1228–1235.

Franke, B., B. M. Neale, and S. V. Faraone, 2009 Genome-wide association studies in ADHD. Hum. Genet. 126: 13–50.

Franke, A., D. P. B. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith *et al.*, 2010 Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. 42: 1118–1125.

Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve *et al.*, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.

Fuentealba, R. A., Q. Liu, J. Zhang, T. Kanekiyo, X. Hu *et al.*, 2010 Low-density lipoprotein receptor-related protein 1 (LRP1) mediates neuronal Abeta42 uptake and lysosomal trafficking. PLoS One 5: e11884.

Gelernter, J., H. R. Kranzler, R. Sherva, L. Almasy, A. I. Herman *et al.*, 2015 Genome-wide association study of nicotine dependence in American populations: identification of novel risk loci in both African-Americans and European-Americans. Biol. Psychiatry 77: 493–503.

Guth, S. I. E., K. Schmidt, A. Hess, and M. Wegner, 2009 Adult-onset degeneration of adipose tissue in mice deficient for the Sox8 transcription factor. J. Lipid Res. 50: 1269–1280.

Hagberg, C. E., A. Falkevall, X. Wang, E. Larsson, J. Huusko *et al.*, 2010 Vascular endothelial growth factor B controls endothelial fatty acid uptake. Nature 464: 917–921.

Hallberg, P., N. Eriksson, L. Ibañez, E. Bondon-Guitton, R. Kreutz *et al.*, 2016 Genetic variants associated with antithyroid drug-induced agranulocytosis: a genome-wide association study in a European population. Lancet Diabetes Endocrinol. 4: 507–516.

Heid, I. M., A. U. Jackson, J. C. Randall, T. W. Winkler, L. Qi *et al.*, 2010 Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat. Genet. 42: 949–960.

Hirschhorn, J. N., and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6: 95–108.

Holden, M., S. Deng, L. Wojnowski, and B. Kulle, 2008 GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics 24: 2784–2785.

Hu, H., K. Eggers, W. Chen, M. Garshasbi, M. M. Motazacker *et al.*, 2011 ST3GAL3 mutations impair the development of higher cognitive functions. Am. J. Hum. Genet. 89: 407–414.

Hu, Y., L. Deng, J. Zhang, X. Fang, P. Mei *et al.*, 2015 A pooling genome-wide association study combining a pathway analysis for typical sporadic Parkinson's disease in the Han population of Chinese mainland. Mol. Neurobiol. 53: 4302–4318.

Huang, D. W., B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord *et al.*, 2007 The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 8: R183.

Jia, P., S. Zheng, J. Long, W. Zheng, and Z. Zhao, 2011 dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics 27: 95–102.

Jiang, D.-K., J. Sun, G. Cao, Y. Liu, D. Lin *et al.*, 2012 Genetic variants in STAT4 and HLA-DQ genes confer risk of hepatitis B virus related hepatocellular carcinoma. Nat. Genet. 45: 72–75.

Jostins, L., S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern *et al.*, 2012 Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491: 119–124.

Kanehisa, M., 1997 A database for post-genome analysis. Trends Genet. 13: 375–376.

Kanehisa, M., S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, 2012 KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40: D109–D114.

Karpe, F., and K. E. Pinnick, 2015 Biology of upper-body and lower-body adipose tissue–link to whole-body phenotypes. Nat. Rev. Endocrinol. 11: 90–100.

Kouri, N., O. A. Ross, B. Dombroski, C. S. Younkin, D. J. Serie *et al.*, 2015 Genome-wide association study of corticobasal degeneration identifies risk variants shared with progressive supranuclear palsy. Nat. Commun. 6: 7247.

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon *et al.*, 2010 Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838.

Lee, M. J., J.-K. Lee, J. W. Choi, C.-S. Lee, J. H. Sim *et al.*, 2012 Interleukin-6 induces S100A9 expression in colonic epithelial cells through STAT3 activation in experimental ulcerative colitis. PLoS One 7: e38801.

Lee, S., L. Miropolsky, and M. Wu, 2015 *Skat: Snp-Set (Sequence) Kernel Association Test*. R package version 1.1.2. Available at: https://CRAN.R-project.org/package=SKAT.

Leiserson, M. D. M., J. V. Eldridge, S. Ramachandran, and B. J. Raphael, 2013 Network analysis of GWAS data. Curr. Opin. Genet. Dev. 23: 602–610.

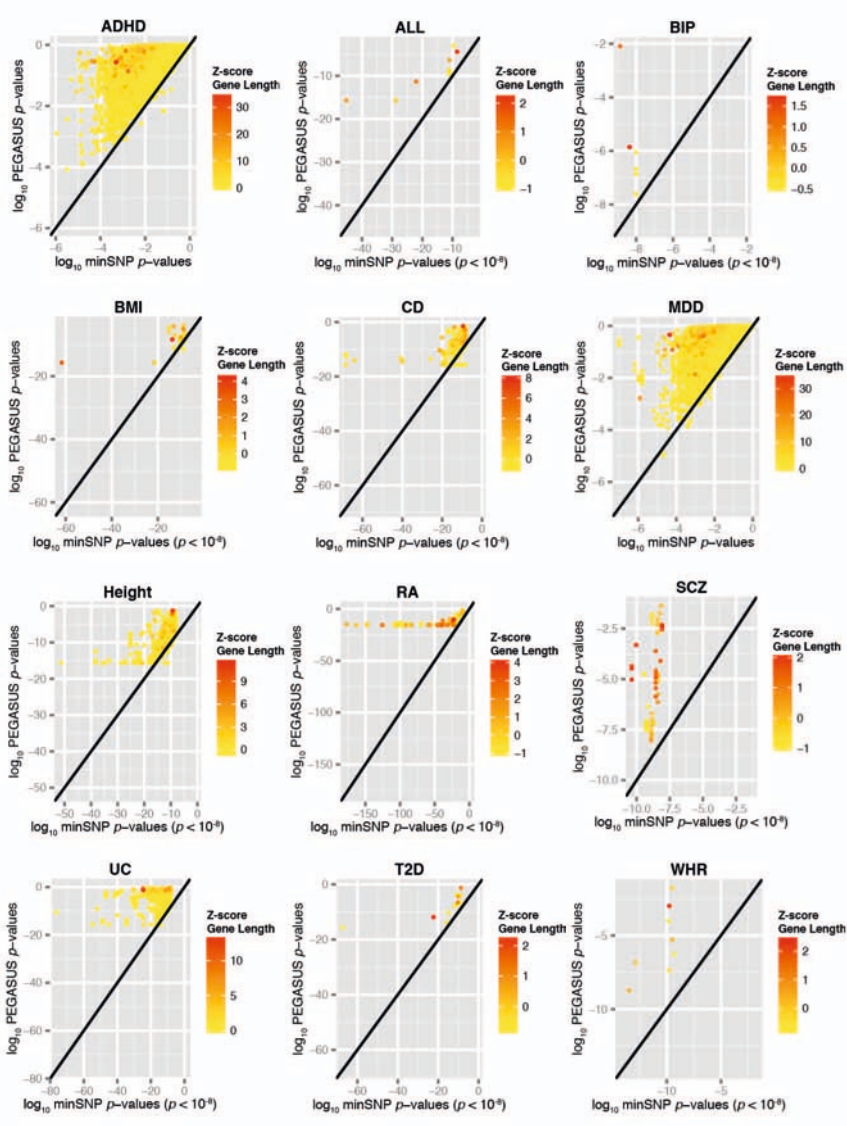
Leiserson, M. D. M., F. Vandin, H.-T. Wu, J. R. Dobson, J. V. Eldridge *et al.*, 2015 Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat. Genet. 47: 106–114.

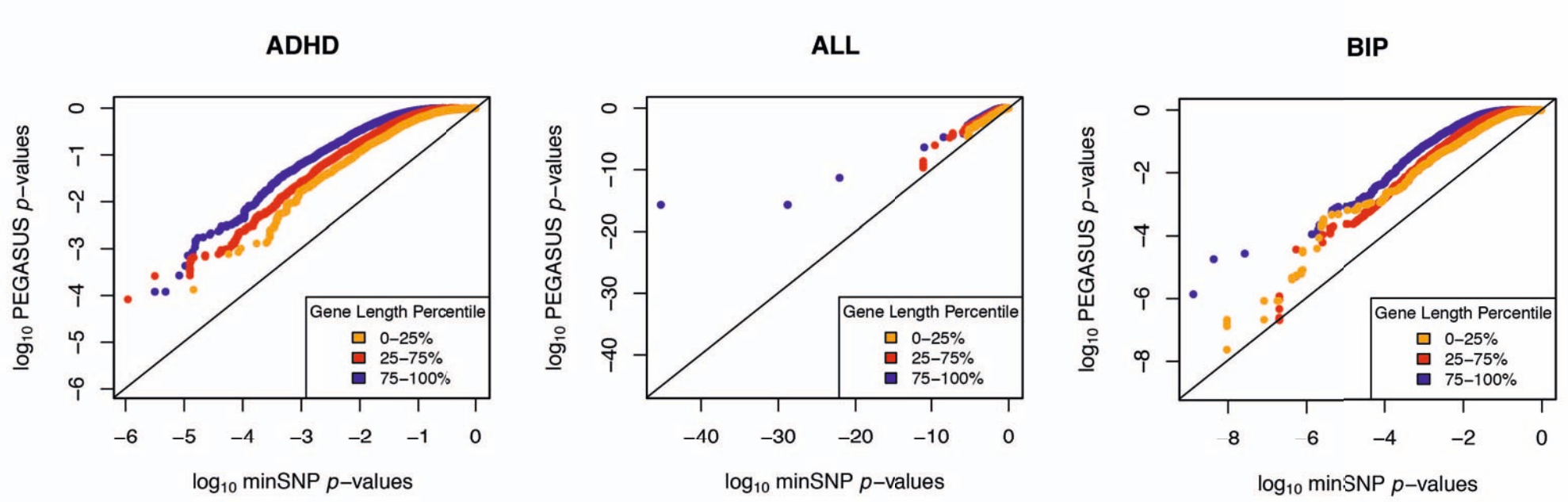
Leslie, R., C. J. O'Donnell, and A. D. Johnson, 2014 GRASP: analysis of genotype-phenotype results from 1390 genome-wide

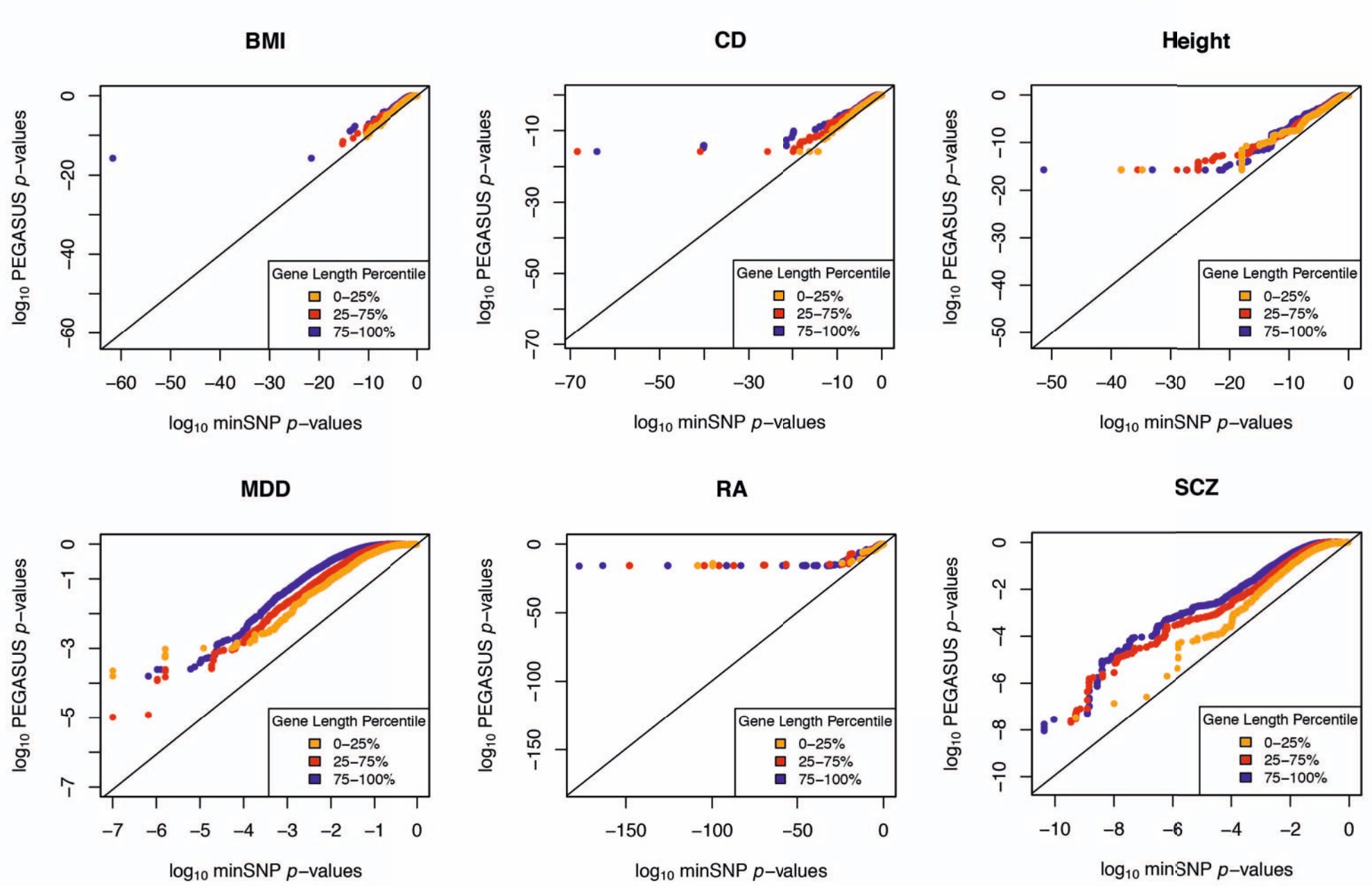
association studies and corresponding open access database. Bioinformatics 30: i185–i194.
Litchfield, K., R. Sultana, A. Renwick, D. Dudakia, S. Seal *et al.*, 2015 Multi-stage genome-wide association study identifies new susceptibility locus for testicular germ cell tumour on chromosome 3q25. Hum. Mol. Genet. 24: 1169–1176.
Liu, D. J., G. M. Peloso, X. Zhan, O. L. Holmen, M. Zawistowski *et al.*, 2013 Meta-analysis of gene-level tests for rare variant association. Nat. Genet. 46: 200–204.
Liu, J. Z., A. F. McRae, D. R. Nyholt, S. E. Medland, N. R. Wray *et al.*, 2010 A versatile gene-based test for genome-wide association studies. Am. J. Hum. Genet. 87: 139–145.
Luciano, M., N. K. Hansell, J. Lahti, G. Davies, S. E. Medland *et al.*, 2011 Whole genome association scan for genetic polymorphisms influencing information processing speed. Biol. Psychol. 86: 193–202.
Maden, M., 2007 Retinoic acid in the development, regeneration and maintenance of the nervous system. Nat. Rev. Neurosci. 8: 755–765.
Masca, N., P. R. Burton, and N. A. Sheehan, 2011 Participant identification in genetic association studies: improved methods and practical implications. Int. J. Epidemiol. 40: 1629–1642.
Mathai, A. M., and S. Provost, 1992 *Quadratic Forms in Random Variables: Theory and Applications*. Marcel Dekker Inc., New York.
McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little *et al.*, 2008 Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9: 356–369.
McClellan, J., and M.-C. King, 2010 Genetic heterogeneity in human disease. Cell 141: 210–217.
McGovern, D. P. B., A. Gardet, L. Törkvist, P. Goyette, J. Essers *et al.*, 2010 Genome-wide association identifies multiple ulcerative colitis susceptibility loci. Nat. Genet. 42: 332–337.
McGrath, C. L., S. J. Glatt, P. Sklar, H. Le-Niculescu, R. Kuczenski *et al.*, 2009 Evidence for genetic association of RORB with bipolar disorder. BMC Psychiatry 9: 70.
Mooney, M. A., J. T. Nigg, S. K. McWeeney, and B. Wilmot, 2014 Functional and genomic context in pathway analysis of GWAS data. Trends Genet. 30: 390–400.
Morris, A. P., B. F. Voight, T. M. Teslovich, T. Ferreira, A. V. Segrè *et al.*, 2012 Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. 44: 981–990.
Naka, H., S. Nakamura, T. Shimazaki, and H. Okano, 2008 Requirement for COUP-TFI and II in the temporal specification of neural stem cells in CNS development. Nat. Neurosci. 11: 1014–1023.
Nalls, M. A., N. Pankratz, C. M. Lill, C. B. Do, D. G. Hernandez *et al.*, 2014 Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. Nat. Genet. 46: 989–993.
Neale, B. M., S. E. Medland, S. Ripke, P. Asherson, B. Franke *et al.*, 2010 Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. J. Am. Acad. Child Adolesc. Psychiatry 49: 884–897.
Need, A. C., D. K. Attix, J. M. McEvoy, E. T. Cirulli, K. L. Linney *et al.*, 2009 A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB. Hum. Mol. Genet. 18: 4650–4661.
Pan, W., 2009 Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet. Epidemiol. 33: 497–507.
Parkes, M., A. Cortes, D. A. van Heel, and M. A. Brown, 2013 Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat. Rev. Genet. 14: 661–673.
Peixoto, L., and T. Abel, 2013 The role of histone acetylation in memory formation and cognitive impairments. Neuropsychopharmacology 38: 62–76.
Peloso, G. M., P. L. Auer, J. C. Bis, A. Voorman, A. C. Morrison *et al.*, 2014 Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am. J. Hum. Genet. 94: 223–232.
Peng, G., L. Luo, H. Siu, Y. Zhu, P. Hu *et al.*, 2010 Gene and pathway-based second-wave analysis of genome-wide association studies. Eur. J. Hum. Genet. 18: 111–117.
Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904–909.
Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.
Raelson, J. V., R. D. Little, A. Ruether, H. Fournier, B. Paquin *et al.*, 2007 Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. Proc. Natl. Acad. Sci. USA 104: 14747–14752.
Raychaudhuri, S., R. M. Plenge, E. J. Rossin, A. C. Y. Ng, S. M. Purcell *et al.*, 2009 Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 5: e1000534.
Renton, A. E., H. A. Pliner, C. Provenzano, A. Evoli, R. Ricciardi *et al.*, 2015 A genome-wide association study of myasthenia gravis. JAMA Neurol. 72: 396–404.
Rietveld, C. A., S. E. Medland, J. Derringer, J. Yang, T. Esko *et al.*, 2013 GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science 340: 1467–1471.
Ripke, S., A. R. Sanders, K. S. Kendler, D. F. Levinson, P. Sklar *et al.*, 2011 Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. 43: 969–976.
Ripke, S., N. R. Wray, C. M. Lewis, S. P. Hamilton, M. M. Weissman *et al.*, 2013 A mega-analysis of genome-wide association studies for major depressive disorder. Mol. Psychiatry 18: 497–511.
Rivière, J.-B., B. W. M. van Bon, A. Hoischen, S. S. Kholmanskikh, B. J. O'Roak, *et al.*, 2012 De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome. Nat. Genet. 44: 440–444, S1–S2.
Rossin, E. J., K. Lage, S. Raychaudhuri, R. J. Xavier, D. Tatar *et al.*, 2011 Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet. 7: e1001273.
Sartor, R. B., 2006 Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. Nat. Clin. Pract. Gastroenterol. Hepatol. 3: 390–407.
Scharl, M., K. A. Wojtal, H. M. Becker, A. Fischbeck, P. Frei *et al.*, 2012 Protein tyrosine phosphatase nonreceptor type 2 regulates autophagosome formation in human intestinal cells. Inflamm. Bowel Dis. 18: 1287–1302.
Scheid, S., and R. Spang, 2005 twilight; a Bioconductor package for estimating the local false discovery rate. Bioinformatics 21: 2921–2922.
Schuurs-Hoeijmakers, J. H. M., E. C. Oh, L. E. L. M. Vissers, M. E. M. Swinkels, C. Gilissen *et al.*, 2012 Recurrent de novo mutations in PACS1 cause defective cranial-neural-crest migration and define a recognizable intellectual-disability syndrome. Am. J. Hum. Genet. 91: 1122–1127.
Segrè, A. V., L. Groop, V. K. Mootha, M. J. Daly, and D. Altshuler, 2010 Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. 6: e1001058.
Shungin, D., T. W. Winkler, D. C. Croteau-Chonka, T. Ferreira, A. E. Locke *et al.*, 2015 New genetic loci link adipose and insulin biology to body fat distribution. Nature 518: 187–196.
Siervo, M., D. Ruggiero, R. Sorice, T. Nutile, M. Aversano *et al.*, 2012 Body mass index is directly associated with biomarkers

of angiogenesis and inflammation in children and adolescents. Nutrition 28: 262–266.
Silverberg, M. S., J. H. Cho, J. D. Rioux, D. P. B. McGovern, J. Wu *et al.*, 2009 Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. Nat. Genet. 41: 216–220.
Skibola, C. F., S. I. Berndt, J. Vijai, L. Conde, Z. Wang *et al.*, 2014 Genome-wide association study identifies five susceptibility loci for follicular lymphoma outside the HLA region. Am. J. Hum. Genet. 95: 462–471.
Sklar, P., S. Ripke, L. Scott, O. Andreassen, S. Cichon *et al.*, 2011 Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat. Genet. 43: 977–983.
Speliotes, E. K., C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson *et al.*, 2010 Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat. Genet. 42: 937–948.
Stahl, E. A., S. Raychaudhuri, E. F. Remmers, G. Xie, S. Eyre *et al.*, 2010 Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. 42: 508–514.
Stranger, B. E., E. A. Stahl, and T. Raj, 2011 Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187: 367–383.
Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert *et al.*, 2005 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102: 15545–15550.
Tang, B., P. Di Lena, L. Schaffer, S. R. Head, P. Baldi *et al.*, 2011 Genome-wide identification of Bcl11b gene targets reveals role in brain-derived neurotrophic factor signaling. PLoS One 6: e23691.
Tchkonia, T., T. Thomou, Y. Zhu, I. Karagiannides, C. Pothoulakis *et al.*, 2013 Mechanisms and metabolic implications of regional differences among fat depots. Cell Metab. 17: 644–656.
Torkamani, A., E. J. Topol, and N. J. Schork, 2008 Pathway analysis of seven common diseases assessed by genome-wide association. Genomics 92: 265–272.
Tzeng, J.-Y., and D. Zhang, 2007 Haplotype-based association analysis via variance-components score test. Am. J. Hum. Genet. 81: 927–938.
Wan, L., S. S. Molloy, L. Thomas, G. Liu, Y. Xiang *et al.*, 1998 PACS-1 defines a novel gene family of cytosolic sorting proteins required for trans-Golgi network localization. Cell 94: 205–216.
Wang, K., M. Li, and M. Bucan, 2007 Pathway-based approaches for analysis of genomewide association studies. Am. J. Hum. Genet. 81: 1278–1283.
Wang, K., M. Li, and H. Hakonarson, 2010 Analysing biological pathways in genome-wide association studies. Nat. Rev. Genet. 11: 843–854.
Wang, K.-S., X. Liu, Q. Zhang, N. Aragam, and Y. Pan, 2012 Parent-of-origin effects of FAS and PDLIM1 in attention-deficit/hyperactivity disorder. J. Psychiatry Neurosci. 37: 46–52.
Wellcome Trust Case Control Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
Wojcik, G. L., W. H. L. Kao, and P. Duggal, 2015 Relative performance of gene- and pathway-level methods as secondary analyses for genome-wide association studies. BMC Genet. 16: 34.
Woo, D., G. J. Falcone, W. J. Devan, W. M. Brown, A. Biffi *et al.*, 2014 Meta-analysis of genome-wide association studies identifies 1q22 as a susceptibility locus for intracerebral hemorrhage. Am. J. Hum. Genet. 94: 511–521.
Wu, M. C., P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock *et al.*, 2010 Powerful SNP-set analysis for case-control genome-wide association studies. Am. J. Hum. Genet. 86: 929–942.
Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke *et al.*, 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89: 82–93.
Xu, H., W. Yang, V. Perez-Andreu, M. Devidas, Y. Fan *et al.*, 2013 Novel susceptibility variants at 10p12.31–12.2 for childhood acute lymphoblastic leukemia in ethnically diverse populations. J. Natl. Cancer Inst. 105: 733–742.
Zhang, X., J.-i. Hamada, A. Nishimoto, Y. Takahashi, T. Murai *et al.*, 2007 HOXC6 and HOXC11 increase transcription of S100beta gene in BrdU-induced in vitro differentiation of GOTO neuroblastoma cells into Schwannian cells. J. Cell. Mol. Med. 11: 299–306.

Communicating editor: E. Eskin


ADHD
ALL
BIP
BMI
CD
MDD
Height
RA
SCZ
UC
T2D
WHR
log10 PEGASUS p–values
log10 minSNP p–values
log10 minSNP p–values (p < 10-8)
Z-score
Gene Length


ADHD
ALL
BIP
BMI
CD
Height
log10 PEGASUS p−values
log10 minSNP p−values
Gene Length Percentile
0–25%
25–75%
75–100%


MDD
RA
SCZ
T2D
UC
WHR
log10 PEGASUS p−values
log10 minSNP p−values
Gene Length Percentile
0–25%
25–75%
75–100%

ADHD

ALL

BIP

BMI

CD

Height

MDD

RA

SCZ

T2D

UC

WHR

$\log_{10}$ PEGASUS p−values

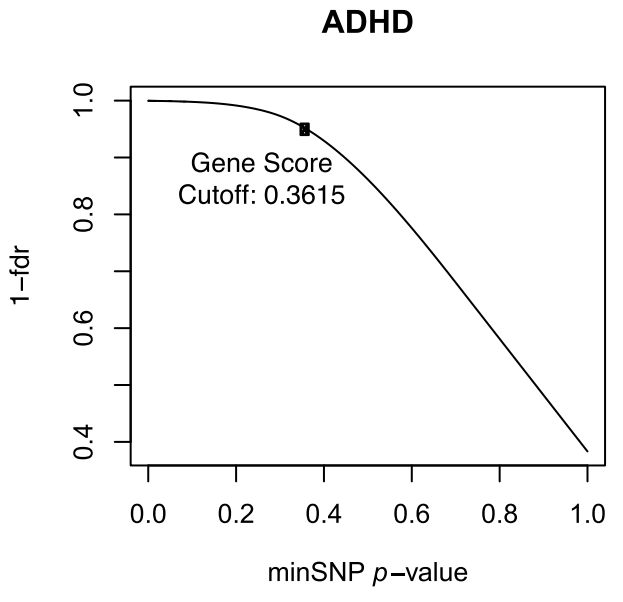
$\log_{10}$ VEGAS p−values

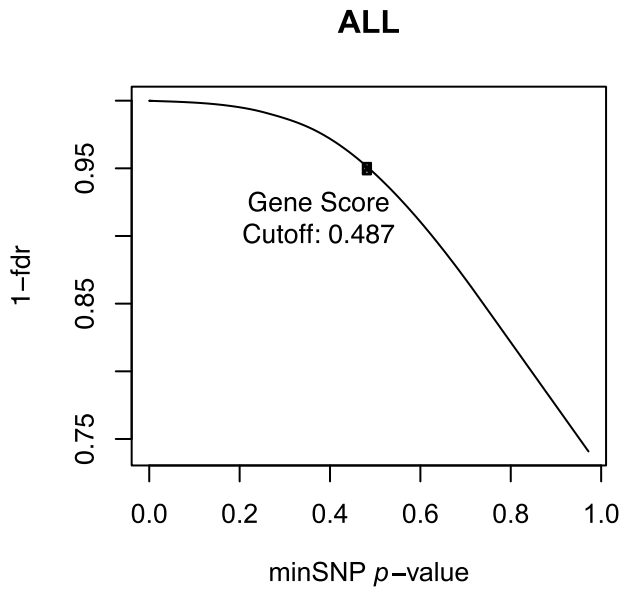
0

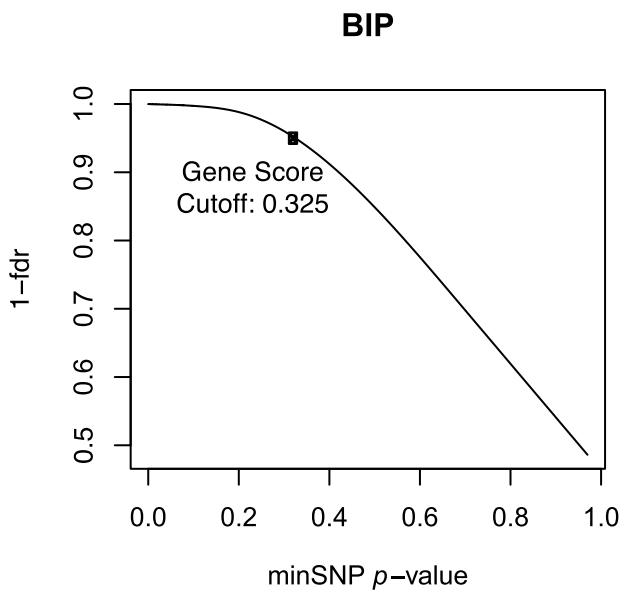
−5

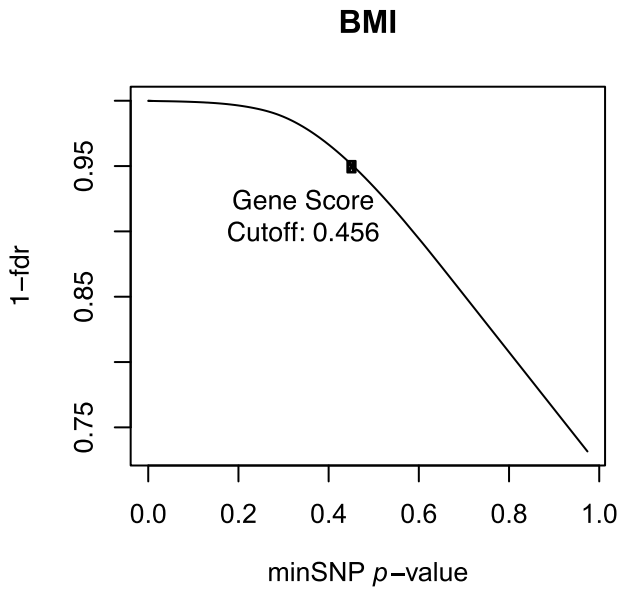
−10

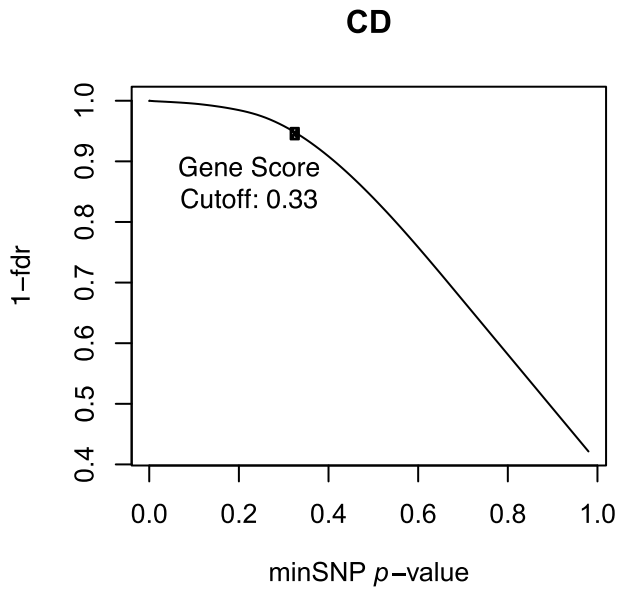
−15

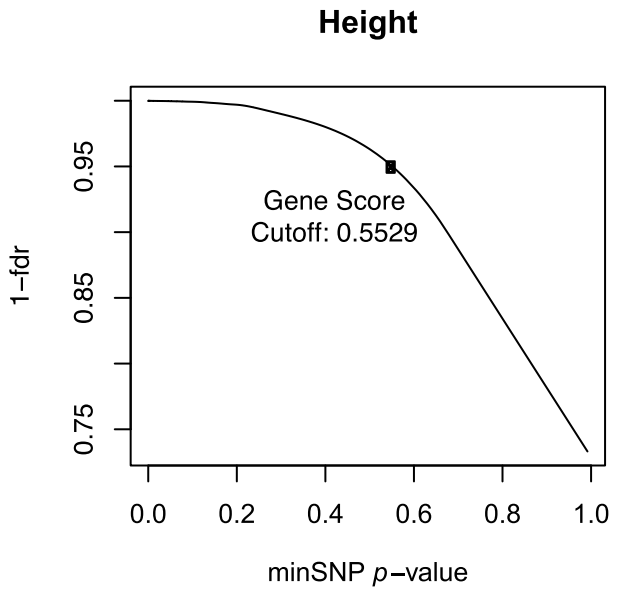

ADHD
Gene Score
Cutoff: 0.3615
1−fdr
minSNP p−value

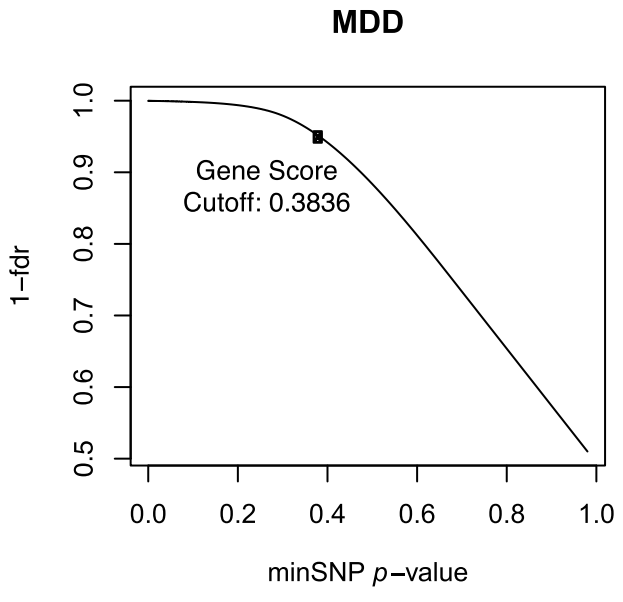

ALL
Gene Score
Cutoff: 0.487
1−fdr
minSNP p−value

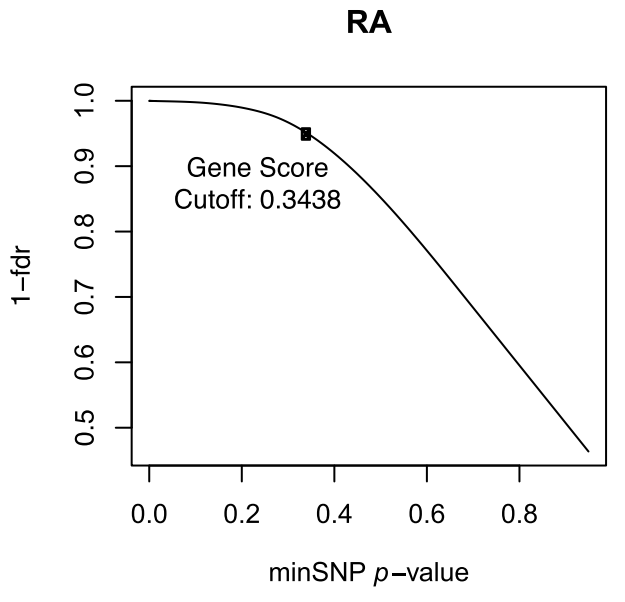

BIP
Gene Score
Cutoff: 0.325
1−fdr
minSNP p−value

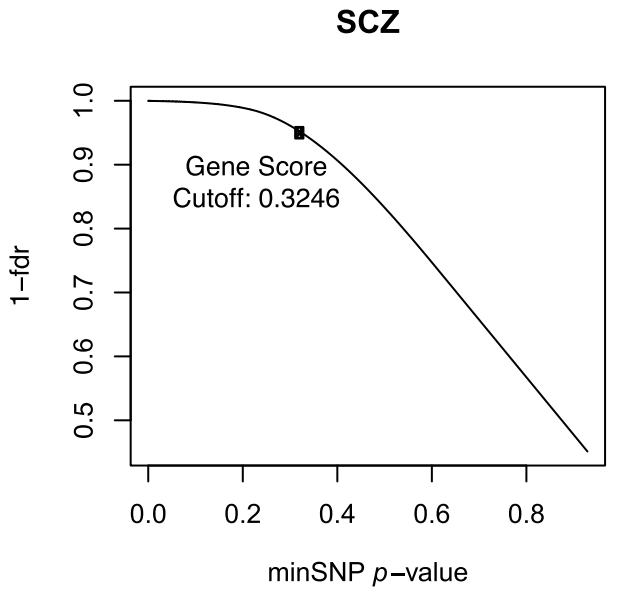

BMI
Gene Score
Cutoff: 0.456
1−fdr
minSNP p−value

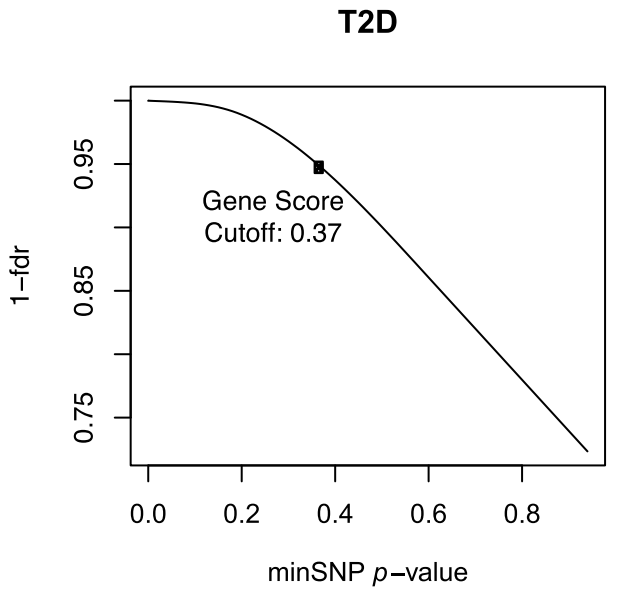

CD
Gene Score
Cutoff: 0.33
1−fdr
minSNP p−value

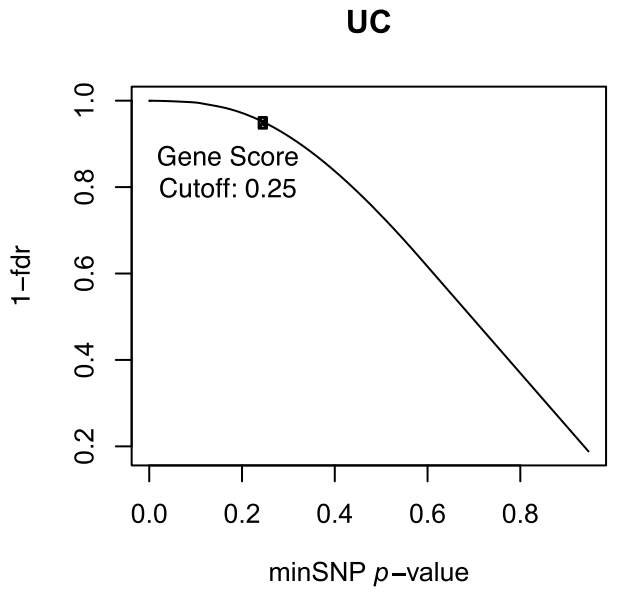

Height
Gene Score
Cutoff: 0.5529
1−fdr
minSNP p−value

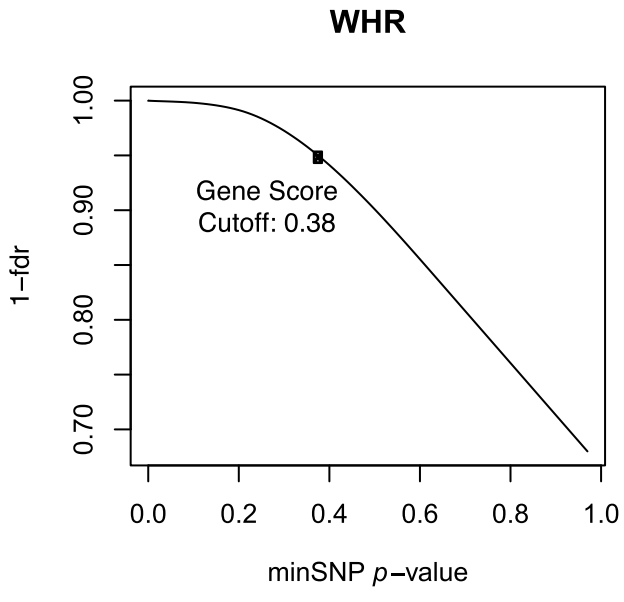

MDD
Gene Score
Cutoff: 0.3836
1−fdr
minSNP p−value

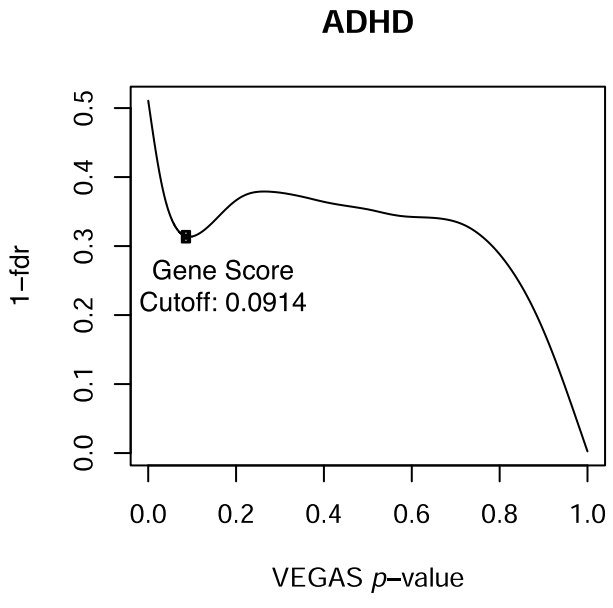

RA
Gene Score
Cutoff: 0.3438
1−fdr
minSNP p−value

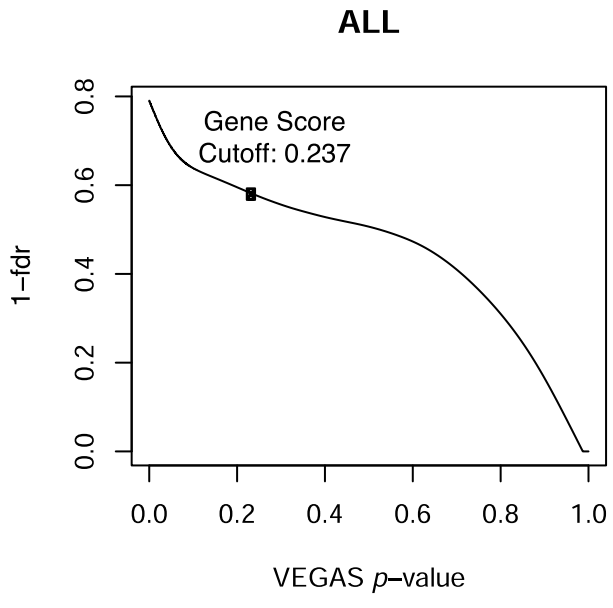

SCZ
Gene Score
Cutoff: 0.3246
1−fdr
minSNP p−value

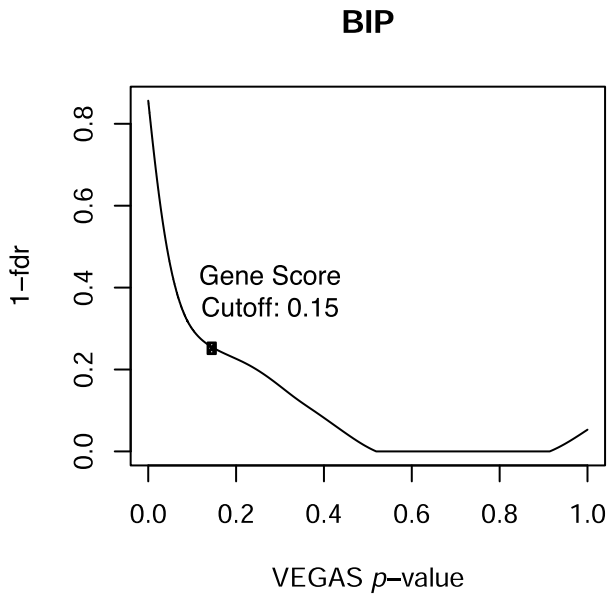

T2D
Gene Score
Cutoff: 0.37
1−fdr
minSNP p−value

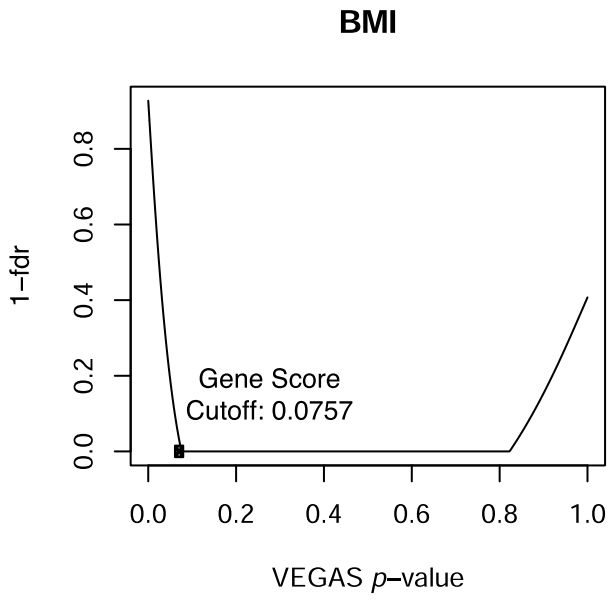

UC
Gene Score
Cutoff: 0.25
1−fdr
minSNP p−value

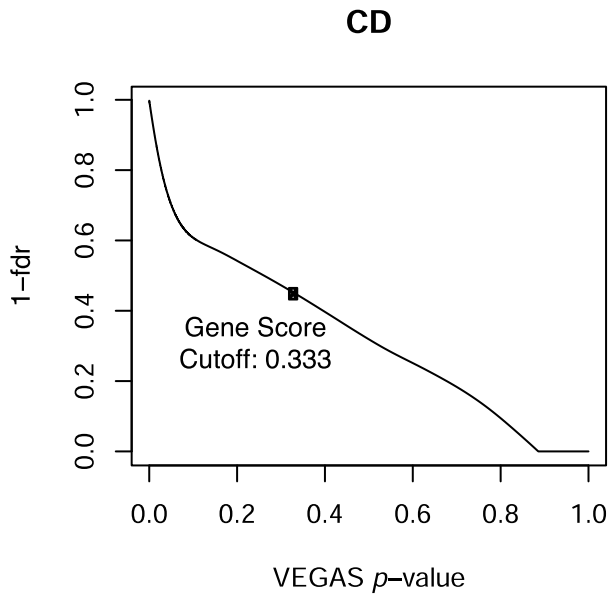

WHR
Gene Score
Cutoff: 0.38
1−fdr
minSNP p−value

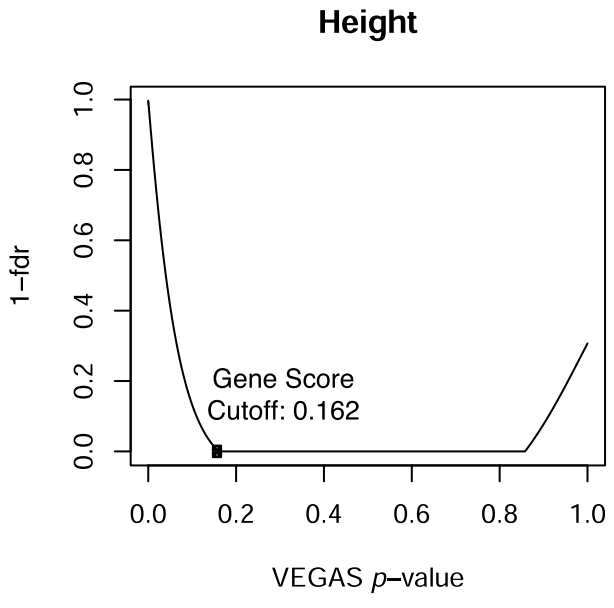

ADHD
1−fdr
Gene Score
Cutoff: 0.0914
VEGAS p−value

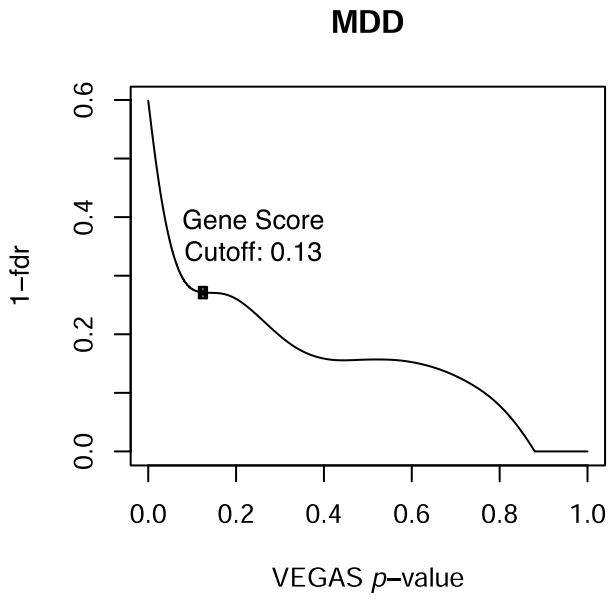

ALL
1−fdr
Gene Score
Cutoff: 0.237
VEGAS p−value

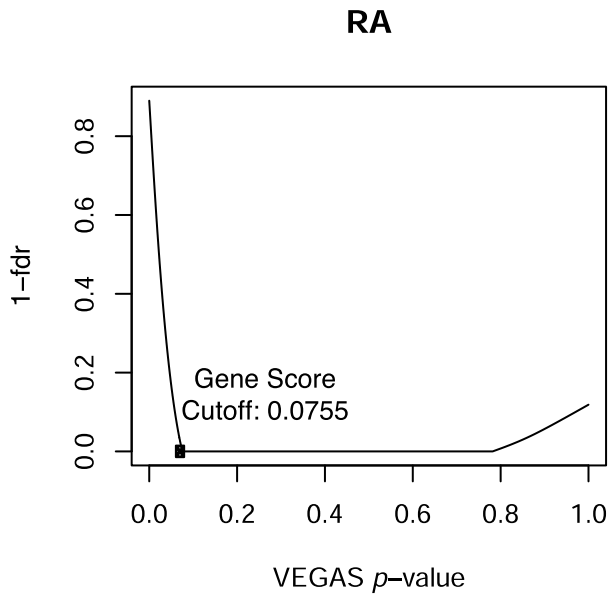

BIP
1−fdr
Gene Score
Cutoff: 0.15
VEGAS p−value

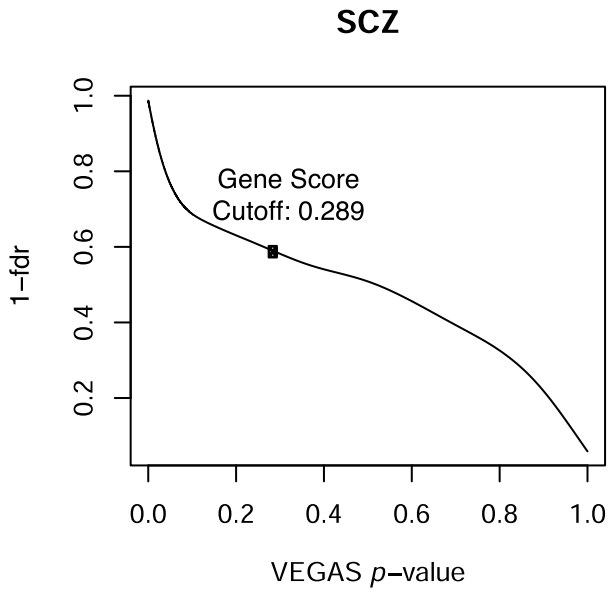

BMI
1−fdr
Gene Score
Cutoff: 0.0757
VEGAS p−value

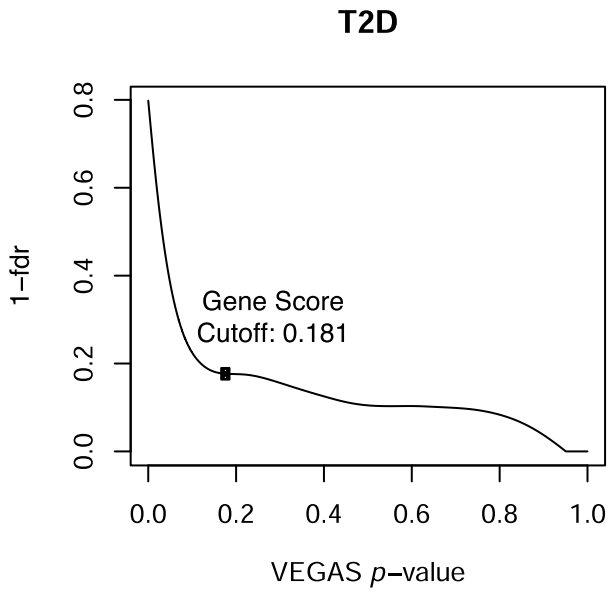

CD
1−fdr
Gene Score
Cutoff: 0.333
VEGAS p−value

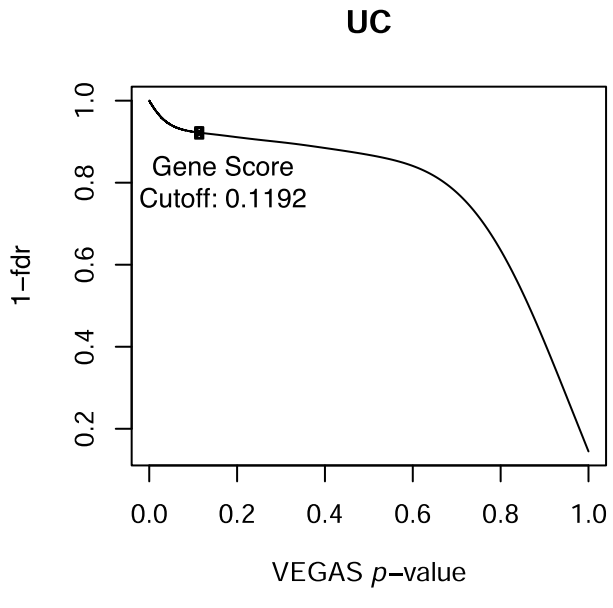

Height
1−fdr
Gene Score
Cutoff: 0.162
VEGAS p−value

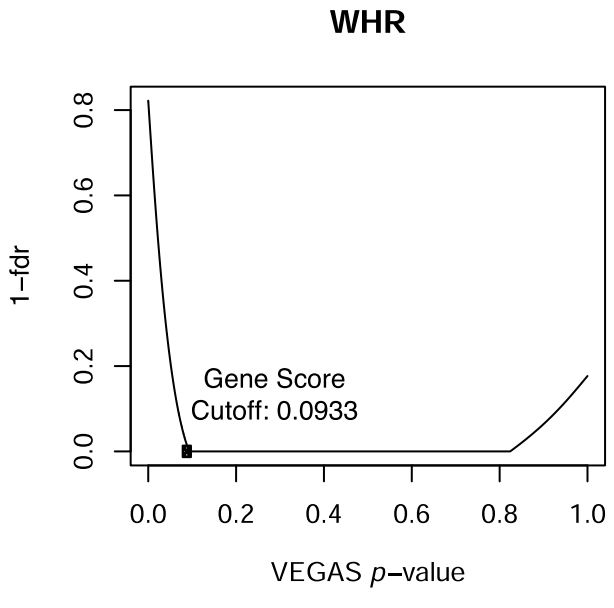

MDD
1−fdr
Gene Score
Cutoff: 0.13
VEGAS p−value

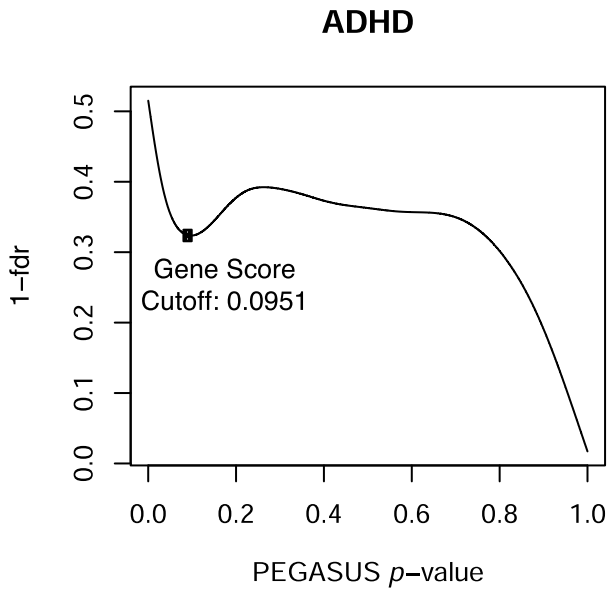

RA
1−fdr
Gene Score
Cutoff: 0.0755
VEGAS p−value

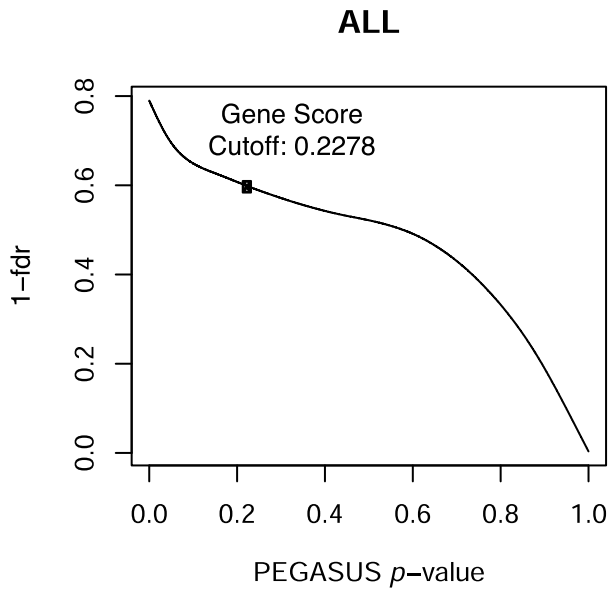

SCZ
1−fdr
Gene Score
Cutoff: 0.289
VEGAS p−value

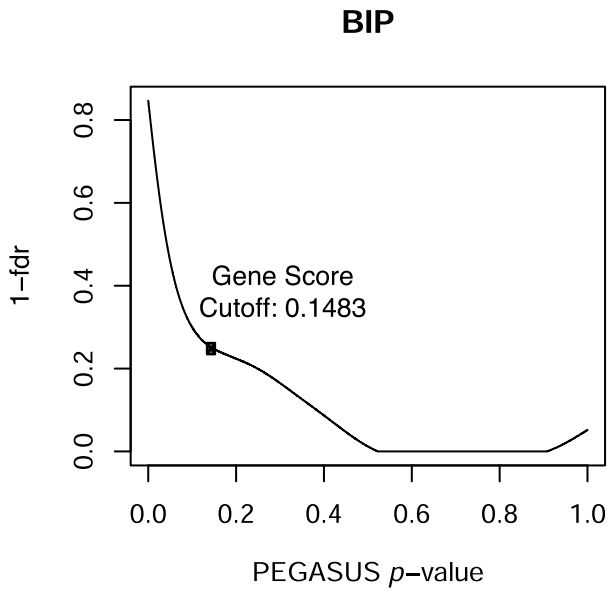

T2D
1−fdr
Gene Score
Cutoff: 0.181
VEGAS p−value

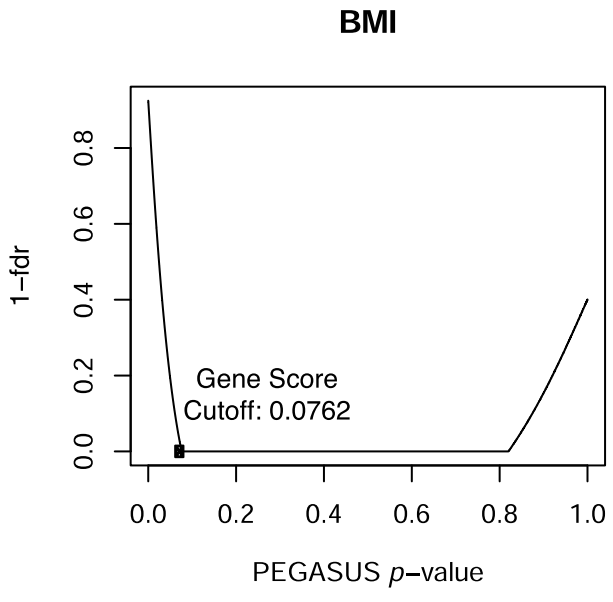

UC
1−fdr
Gene Score
Cutoff: 0.1192
VEGAS p−value

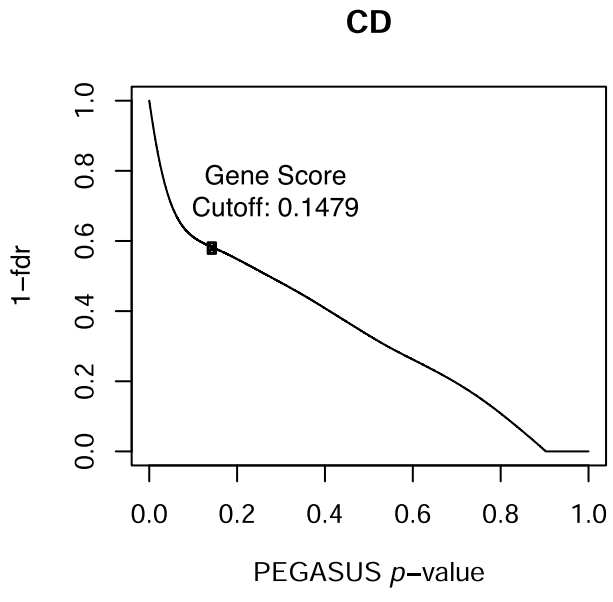

WHR
1−fdr
Gene Score
Cutoff: 0.0933
VEGAS p−value

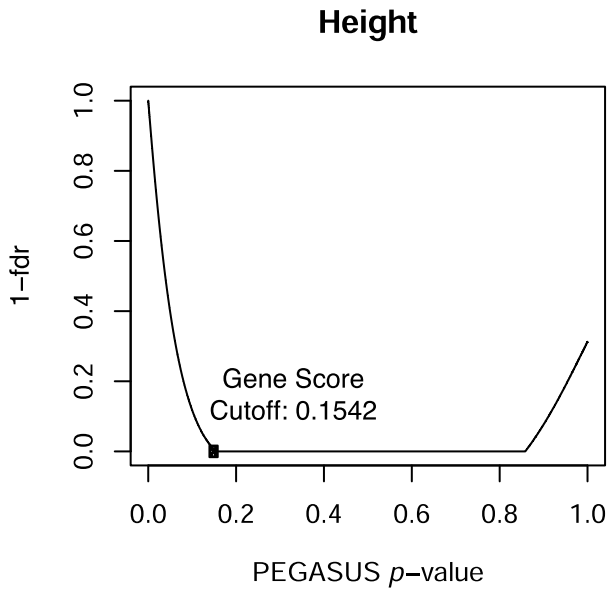

ADHD
Gene Score
Cutoff: 0.0951
1−fdr
PEGASUS p−value

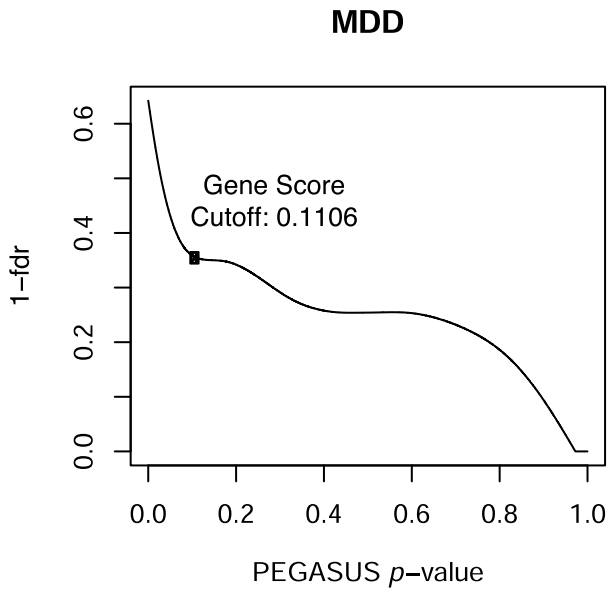

ALL
Gene Score
Cutoff: 0.2278
1−fdr
PEGASUS p−value

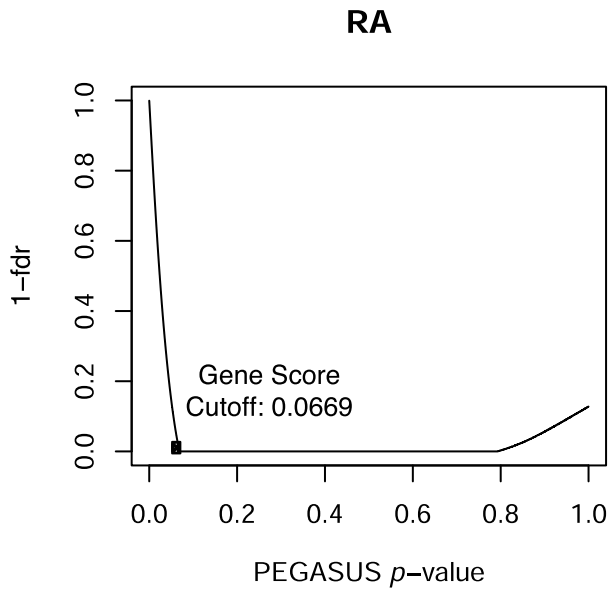

BIP
Gene Score
Cutoff: 0.1483
1−fdr
PEGASUS p−value

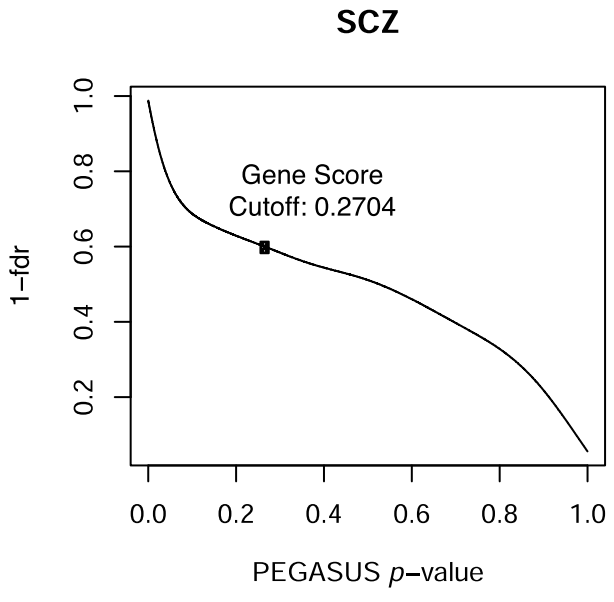

BMI
Gene Score
Cutoff: 0.0762
1−fdr
PEGASUS p−value

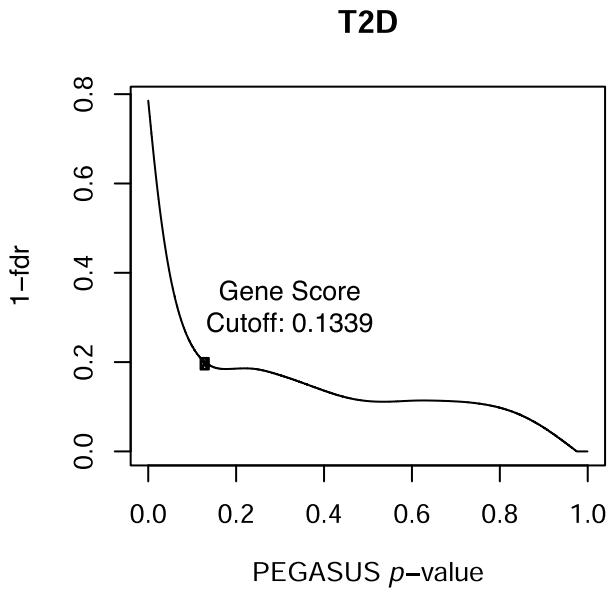

CD
Gene Score
Cutoff: 0.1479
1−fdr
PEGASUS p−value

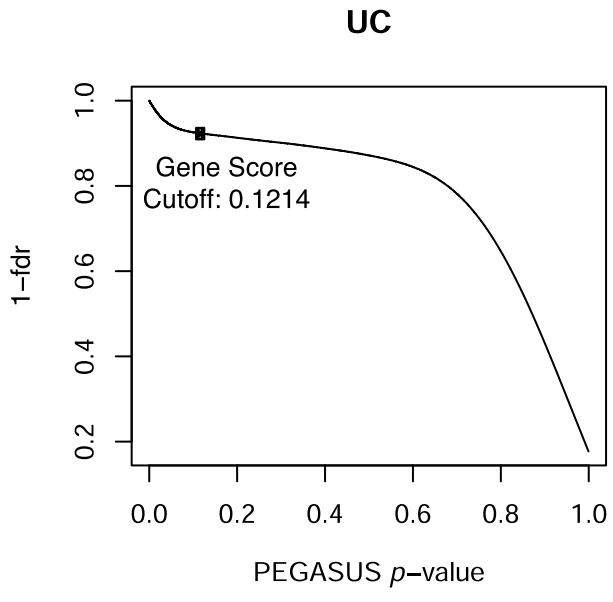

Height
Gene Score
Cutoff: 0.1542
1−fdr
PEGASUS p−value

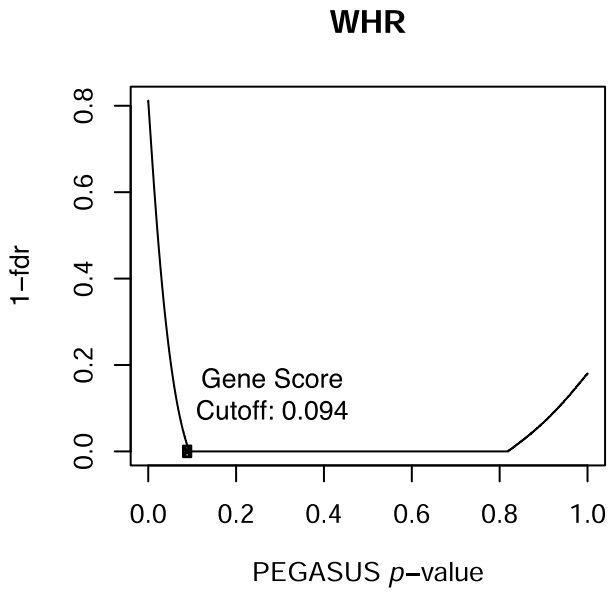

MDD
Gene Score
Cutoff: 0.1106
1−fdr
PEGASUS p−value


RA
Gene Score
Cutoff: 0.0669
1−fdr
PEGASUS p−value

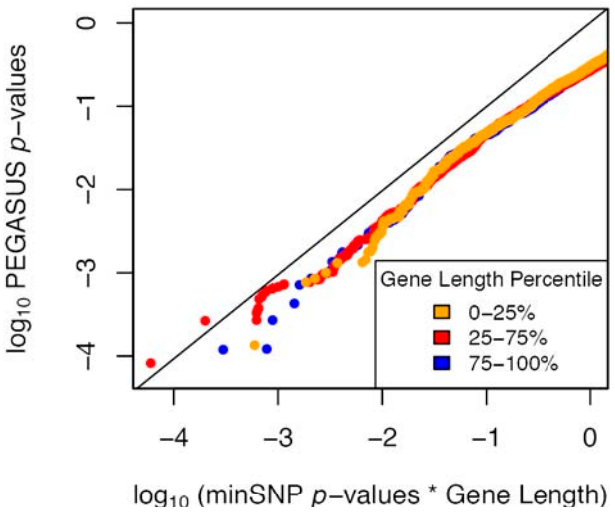

SCZ
Gene Score
Cutoff: 0.2704
1−fdr
PEGASUS p−value


T2D
Gene Score
Cutoff: 0.1339
1−fdr
PEGASUS p−value

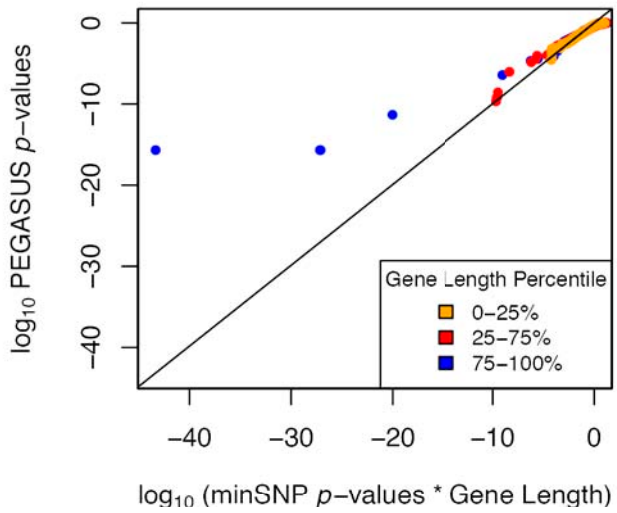

UC
Gene Score
Cutoff: 0.1214
1−fdr
PEGASUS p−value


WHR
Gene Score
Cutoff: 0.094
1−fdr
PEGASUS p−value

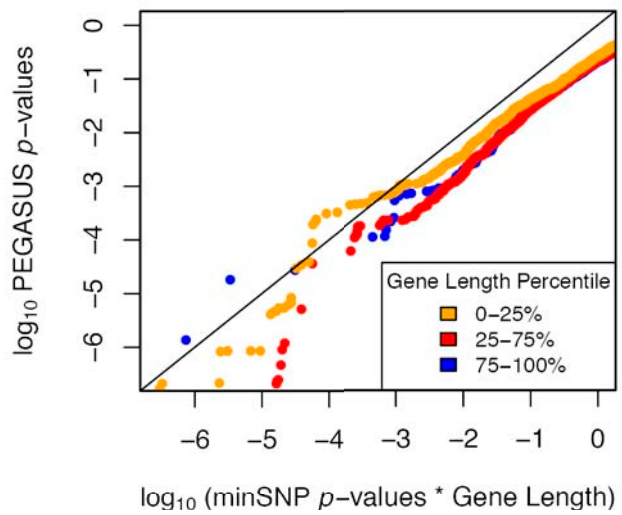
ADHD



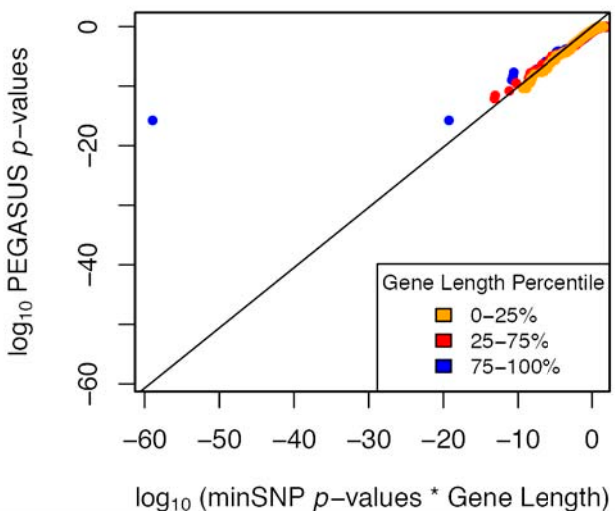
ALL



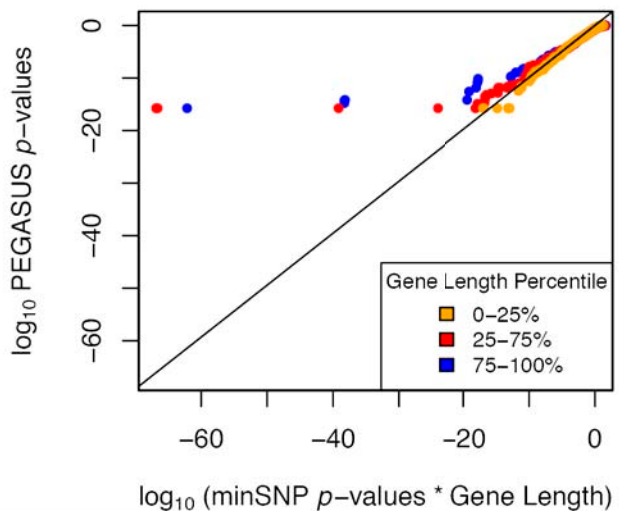
BIP



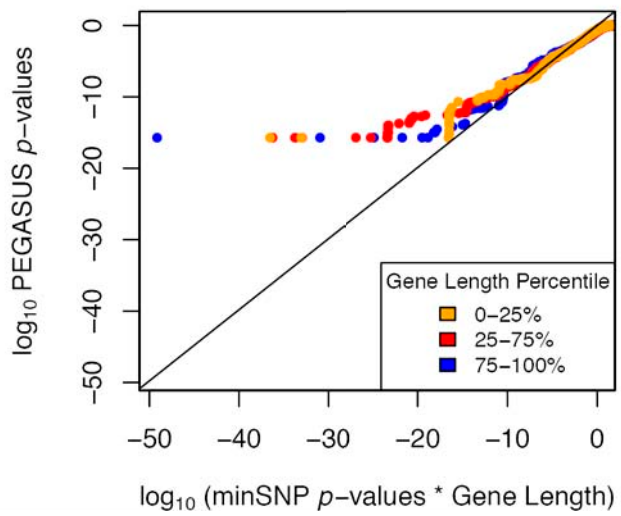
BMI



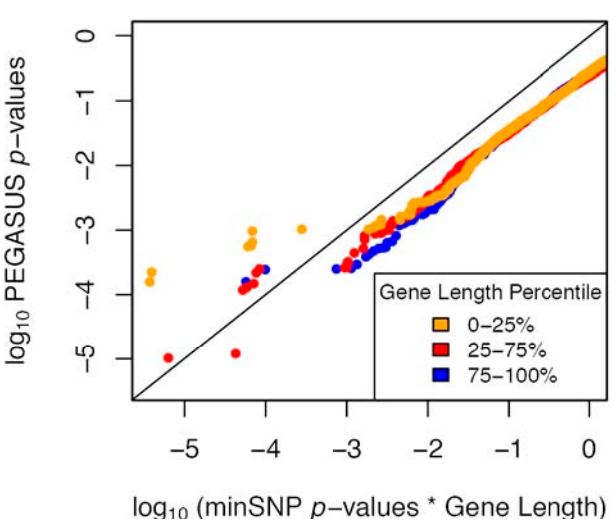
CD



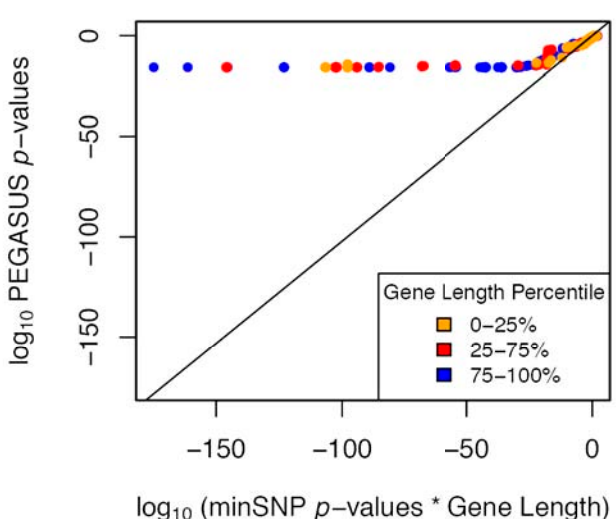
Height



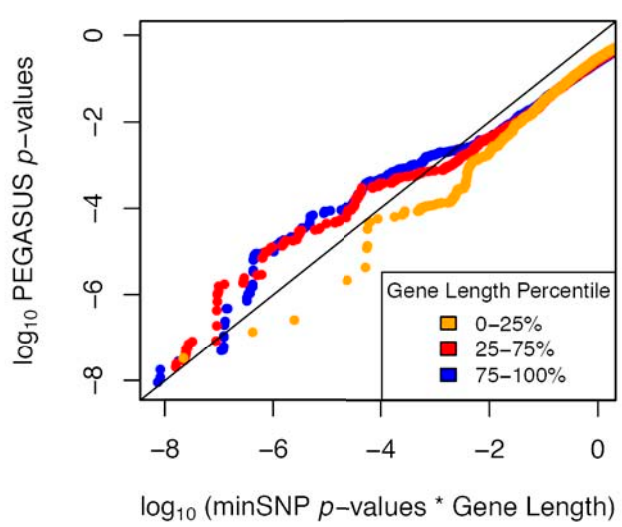
MDD



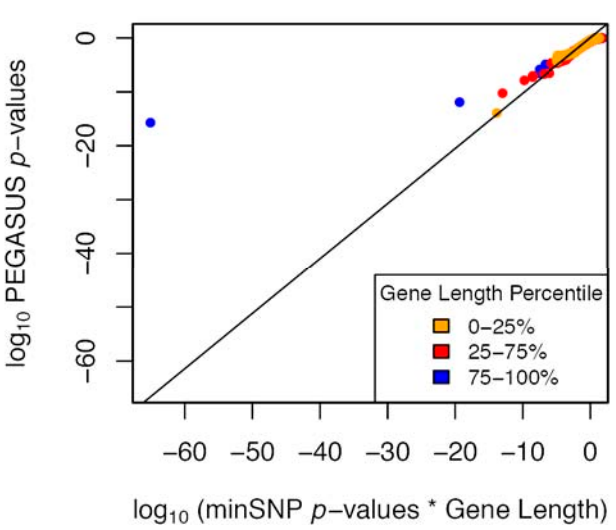
RA



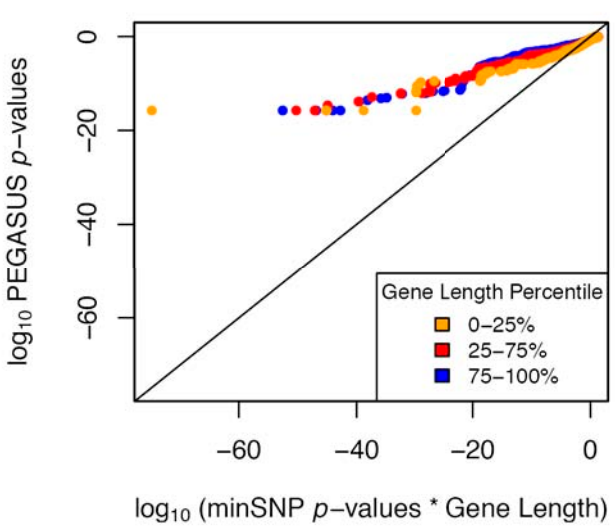
SCZ



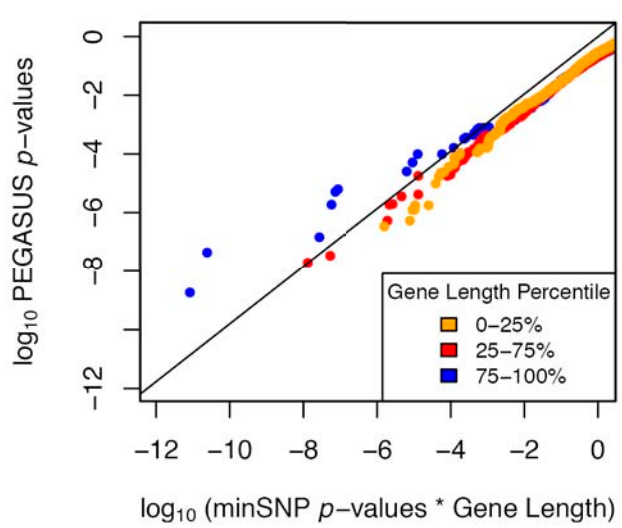
T2D

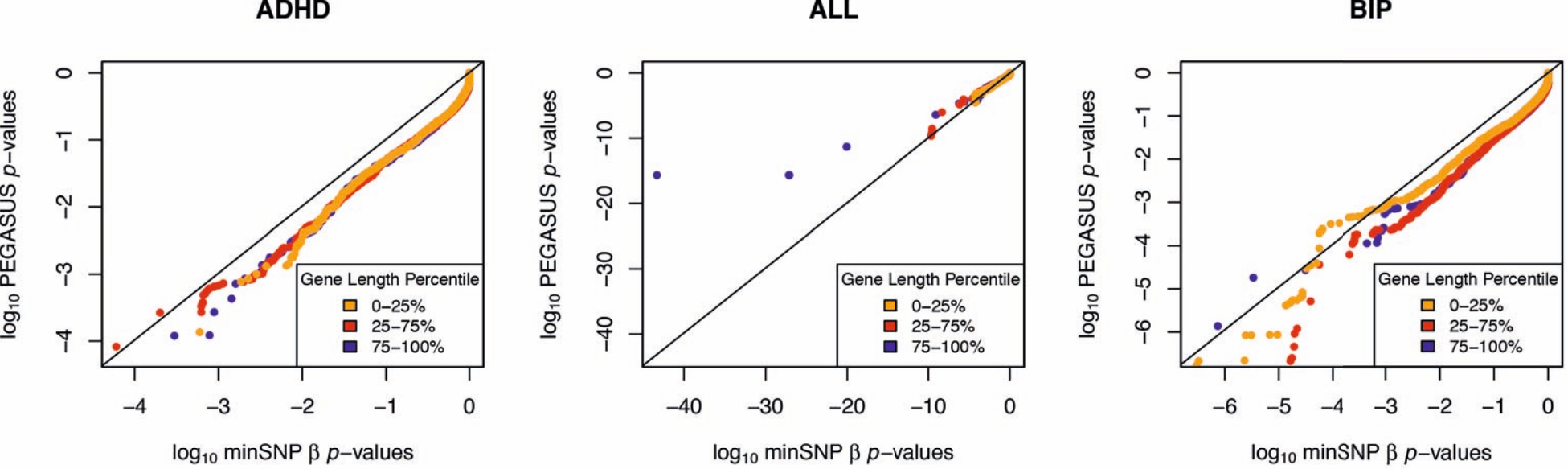


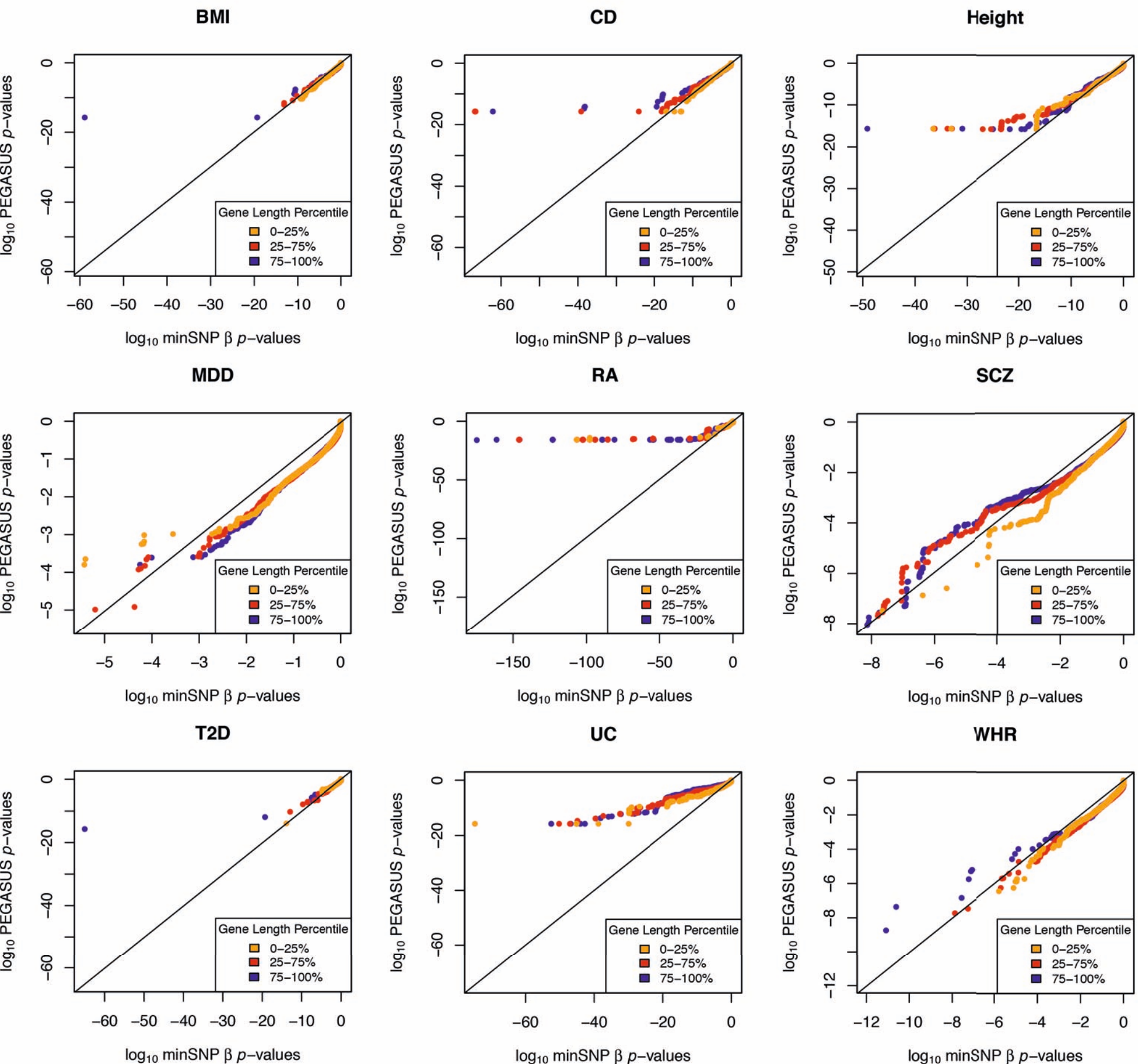
UC

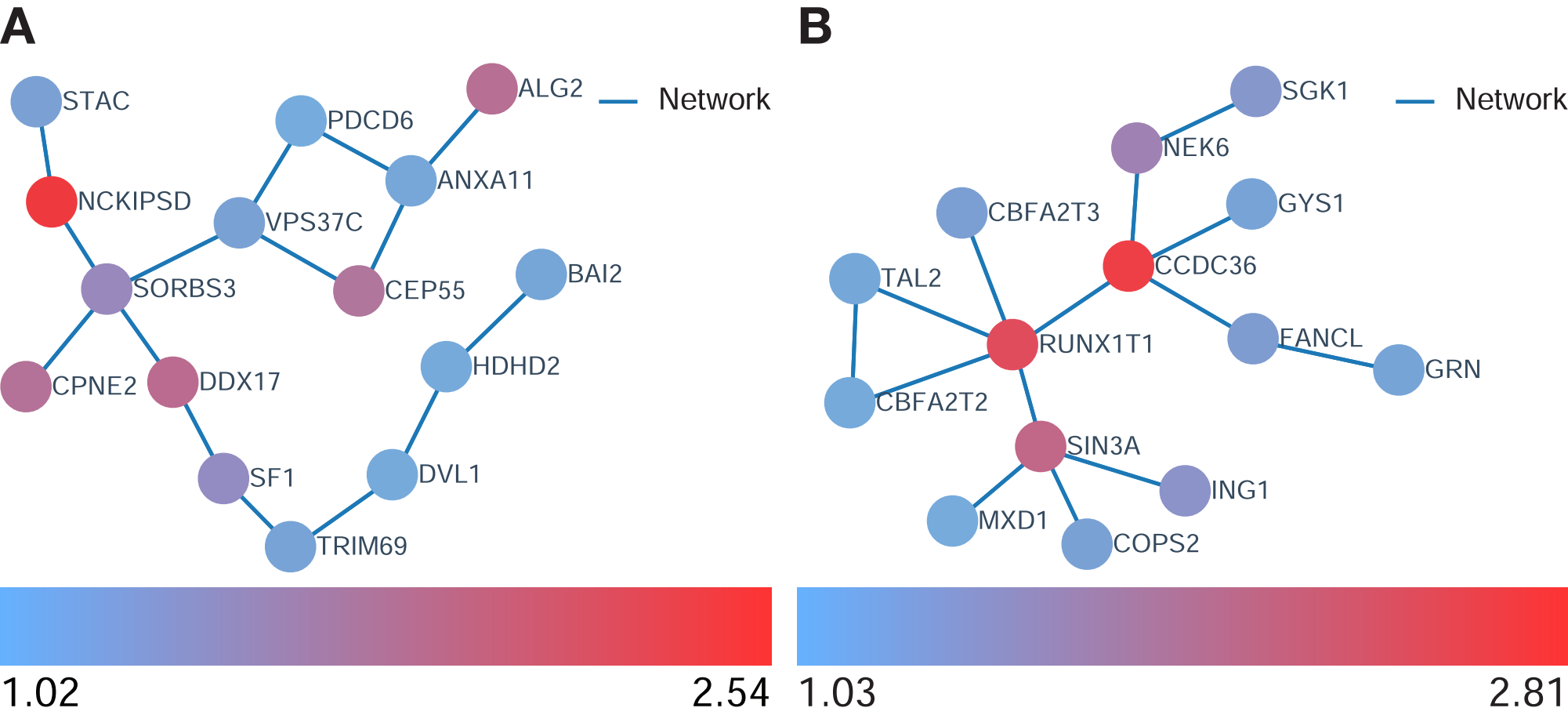


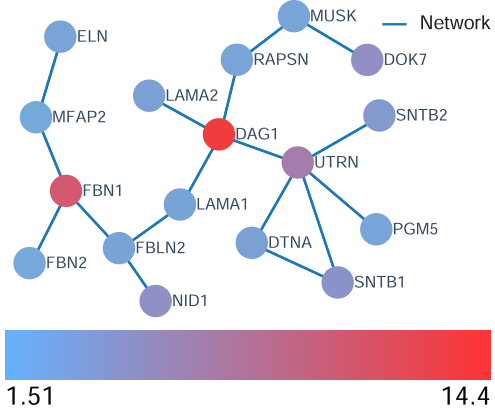
WHR



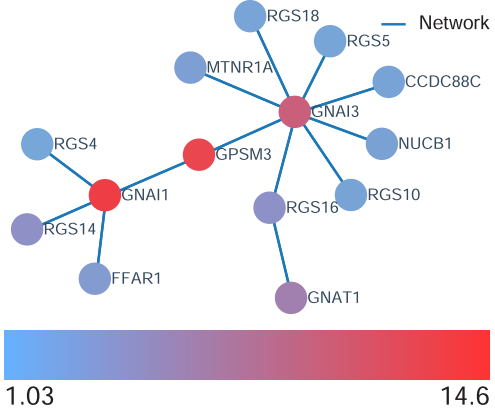

ADHD
ALL
BIP
BMI
CD
Height
log10 PEGASUS p−values
log10 minSNP β p−values
Gene Length Percentile
0–25%
25–75%
75–100%


MDD
RA
SCZ
T2D
UC
WHR
log10 PEGASUS p−values
log10 minSNP β p−values
Gene Length Percentile
0–25%
25–75%
75–100%

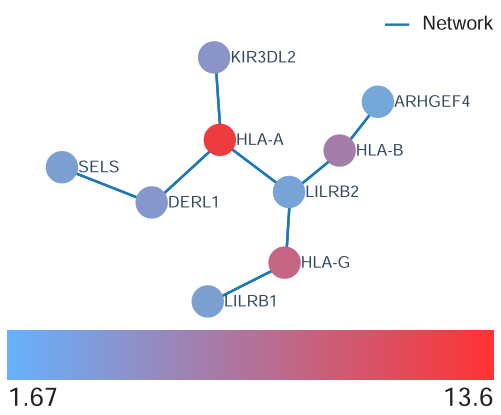

A
STAC
NCKIPSD
SORBS3
VPS37C
PDCD6
ALG2
ANXA11
CEP55
BAI2
HDHD2
CPNE2
DDX17
SF1
DVL1
TRIM69
Network
1.02
2.54
B
SGK1
NEK6
CBFA2T3
GYS1
TAL2
CCDC36
RUNX1T1
FANCL
GRN
CBFA2T2
SIN3A
ING1
MXD1
COPS2
Network
1.03
2.81



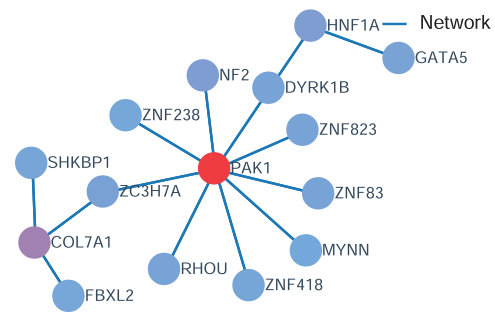
1.51 14.4



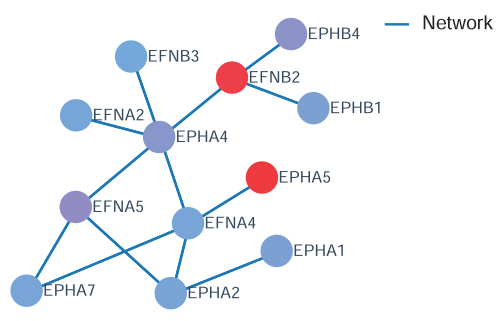
1.03 14.6



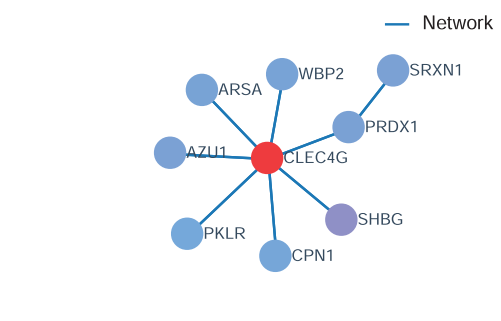
1.67 13.6



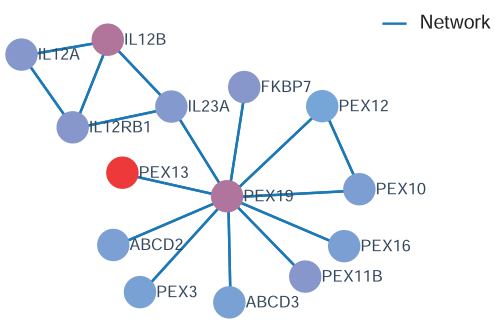
1.18 22.9



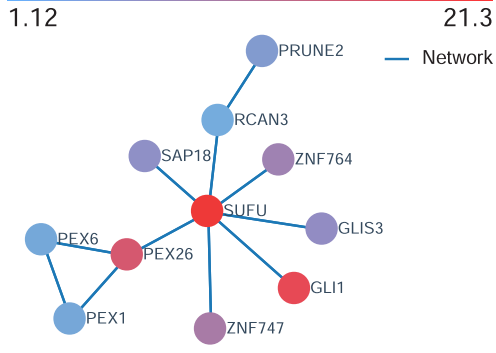
1.12 21.3



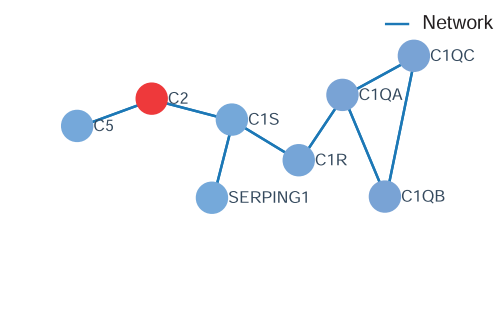
1.14 23.1



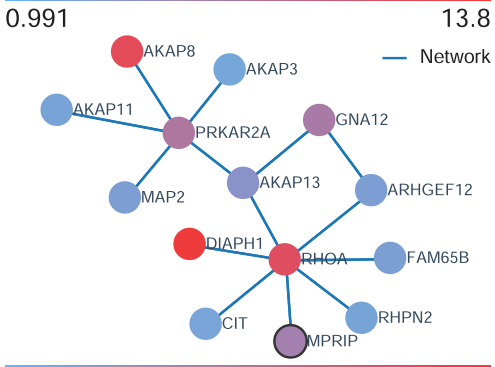
0.991 13.8



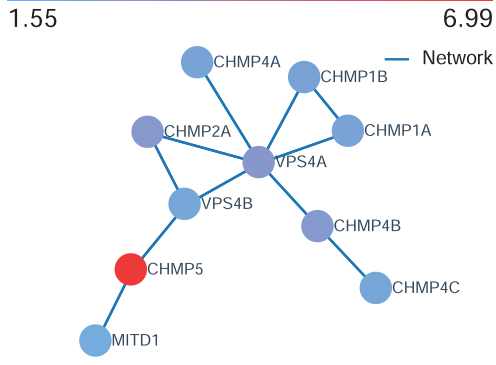
1.55 6.99



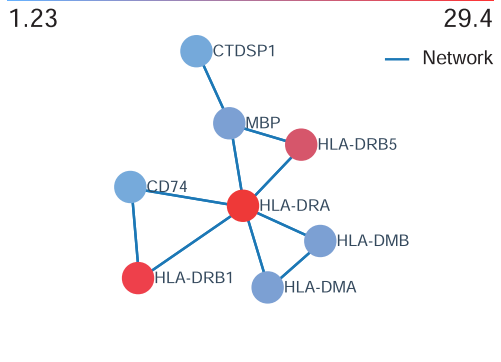
1.23 29.4



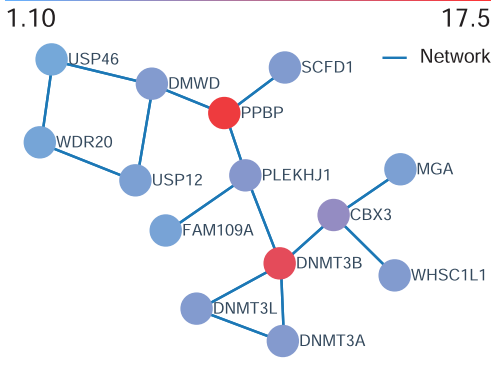
1.10 17.5



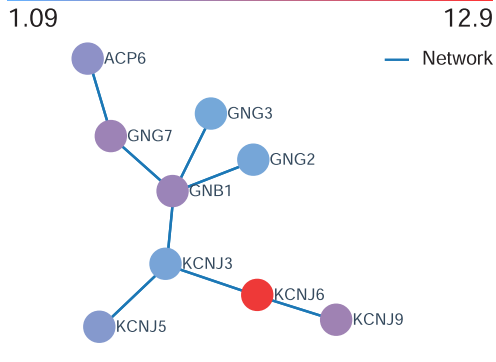
1.09 12.9



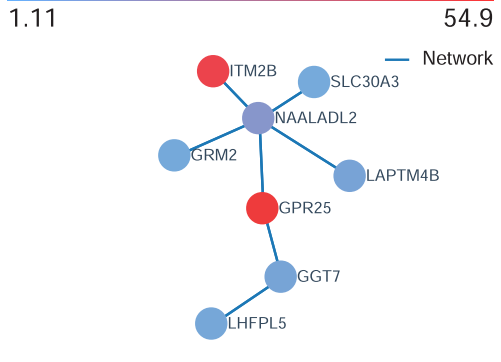
1.11 54.9



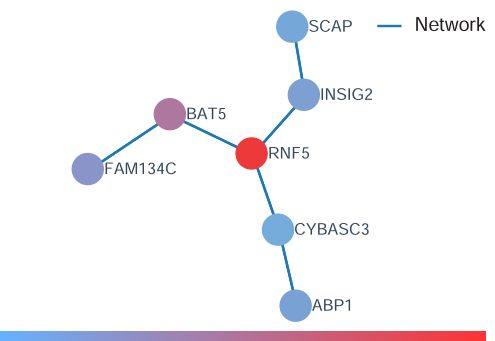
1.06 19.3



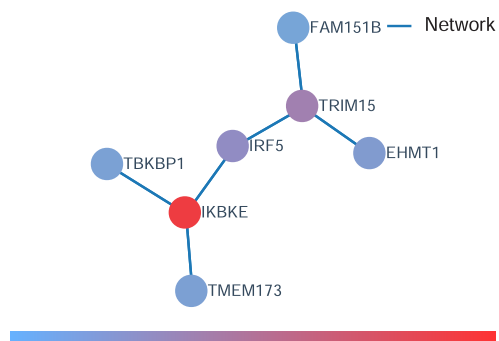
1.78 6.49



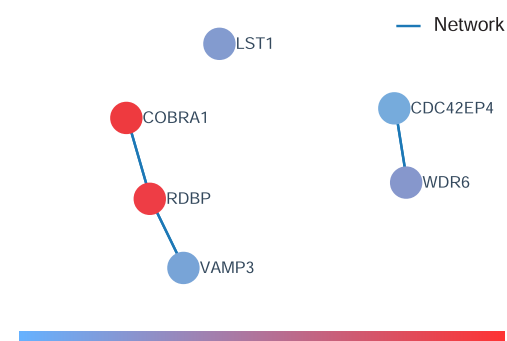
1.25 12.7



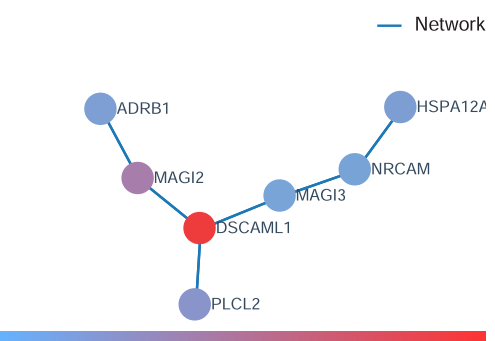
1.09 13.7



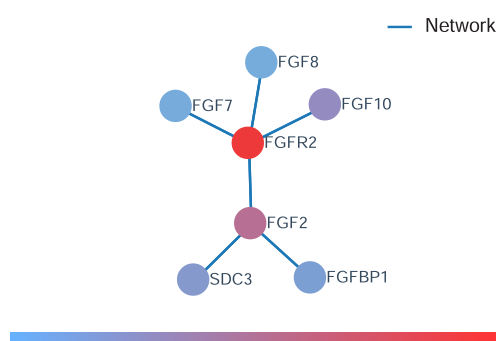
1.50 18.7



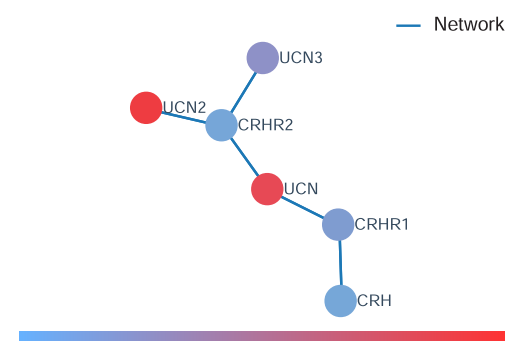
2.73 30.8



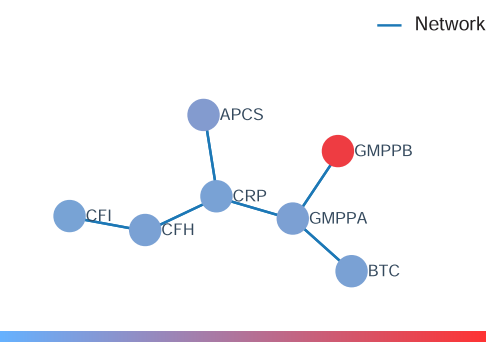
1.68 8.81



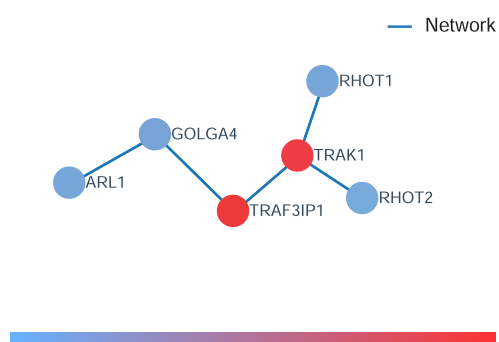
1.03 5.44



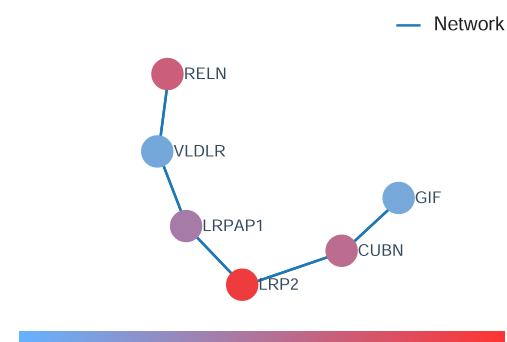
1.28 5.21



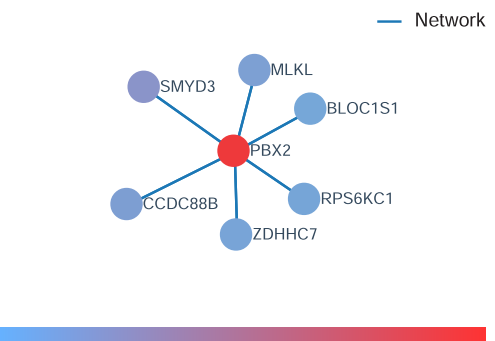
1.80 16.8



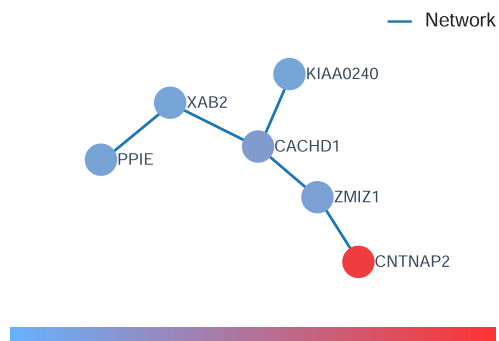
1.19 12.8



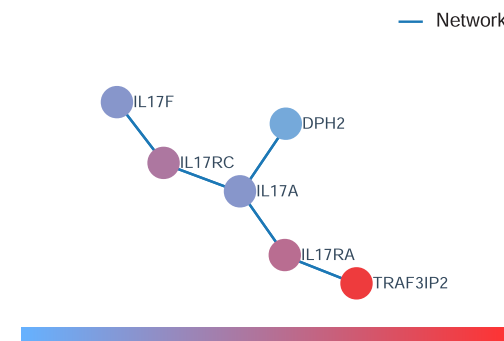
2.24 8.30



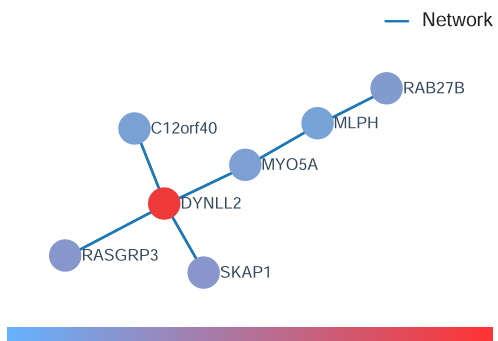
1.87 13.7



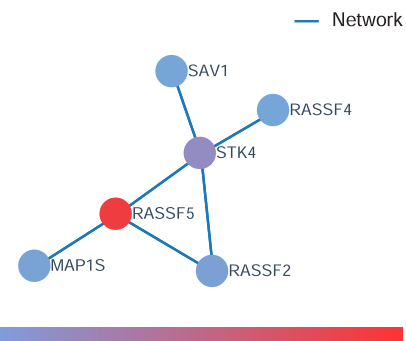
1.94 24.7



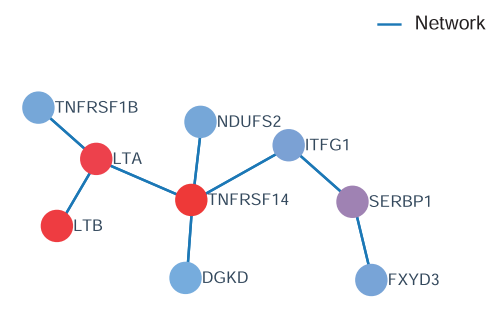
1.21 5.85



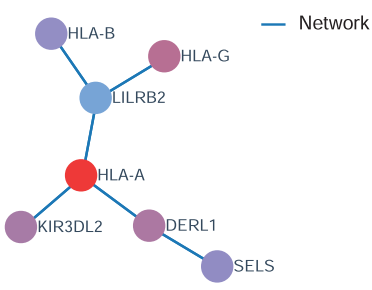
1.54 15.9



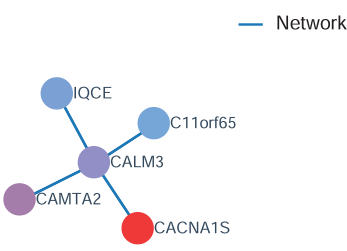
1.62 18.7


Network
TNFRSF1B
NDUFS2
ITFG1
LTA
TNFRSF14
SERBP1
LTB
DGKD
FXYD3

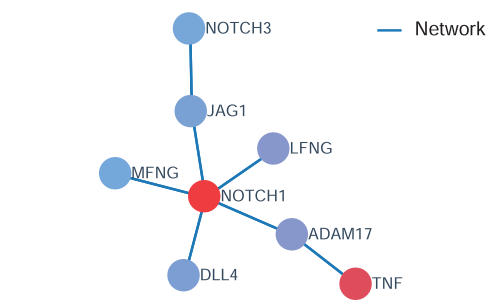

1.06
6.00


Network
HLA-B
HLA-G
LILRB2
HLA-A
KIR3DL2
DERL1
SELS

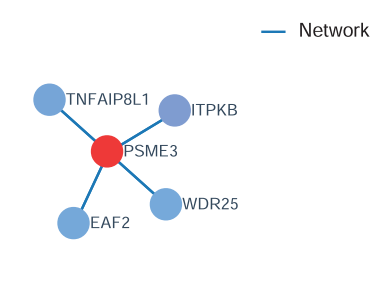

0.974
6.00


Network
IQCE
C11orf65
CALM3
CAMTA2
CACNA1S

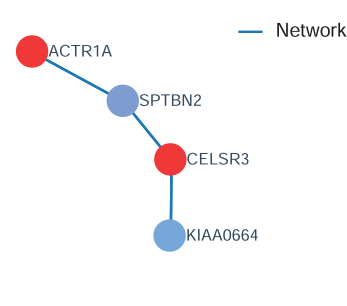

0.924
6.00


Network
NOTCH3
JAG1
LFNG
MFNG
NOTCH1
ADAM17
DLL4
TNF

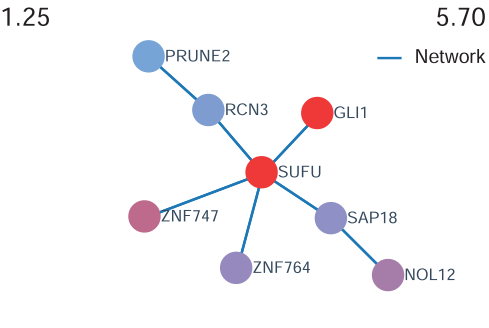

1.25
5.70


Network
TNFAIP8L1
ITPKB
PSME3
WDR25
EAF2

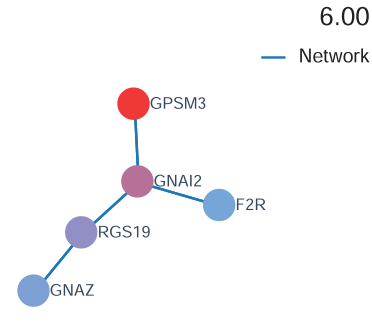

1.18
6.00


Network
ACTR1A
SPTBN2
CELSR3
KIAA0664

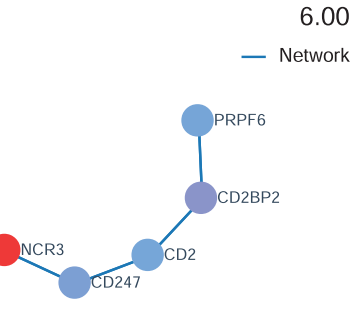

1.27
6.00


Network
PRUNE2
RCN3
GLI1
SUFU
ZNF747
SAP18
ZNF764
NOL12

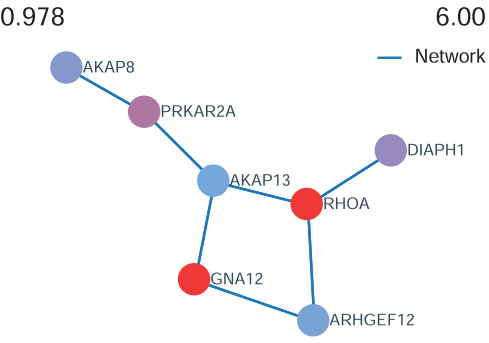

0.978
6.00


Network
GPSM3
GNAI2
F2R
RGS19
GNAZ

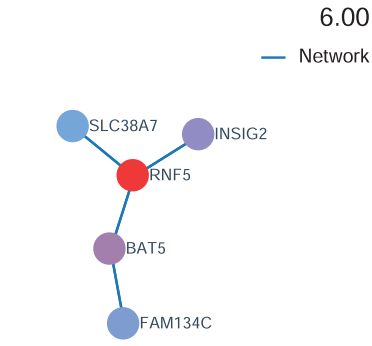

1.39
6.00


Network
PRPF6
CD2BP2
NCR3
CD2
CD247

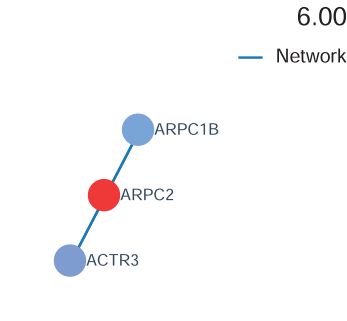

1.36
6.00


Network
AKAP8
PRKAR2A
DIAPH1
AKAP13
RHOA
GNA12
ARHGEF12

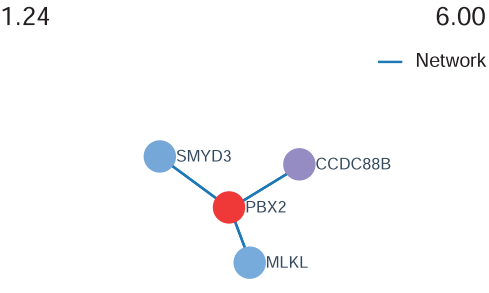

1.24
6.00

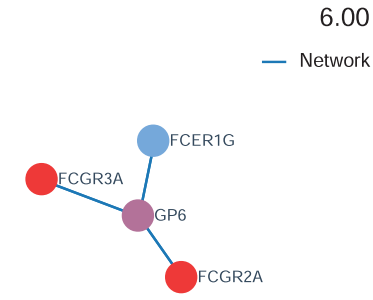

Network
SLC38A7
INSIG2
RNF5
BAT5
FAM134C

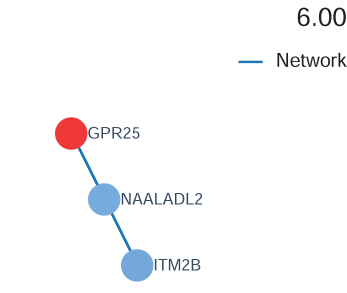

0.921
6.00


Network
ARPC1B
ARPC2
ACTR3

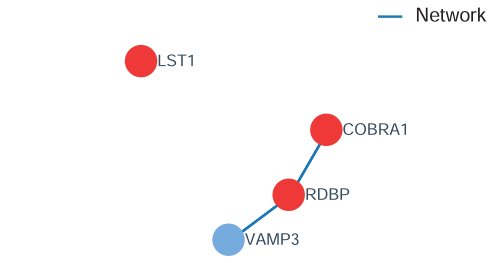

0.971
6.00


Network
SMYD3
CCDC88B
PBX2
MLKL

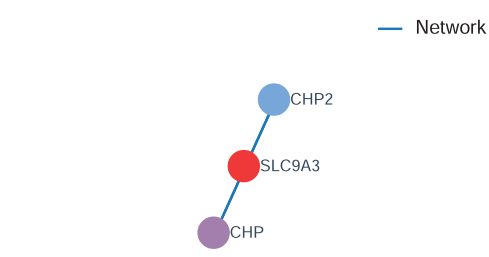

Network
FCER1G
FCGR3A
GP6
FCGR2A


Network
GPR25
NAALADL2
ITM2B

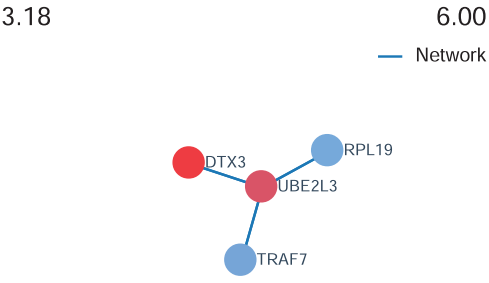

1.06
6.00


Network
LST1
COBRA1
RDBP
VAMP3


2.03
6.00

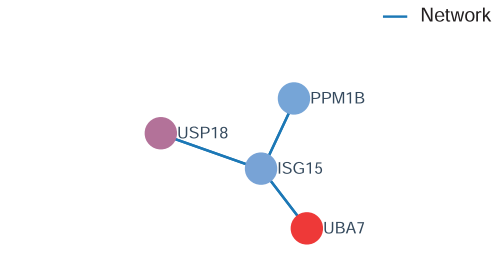

Network
CHP2
SLC9A3
CHP


3.18
6.00

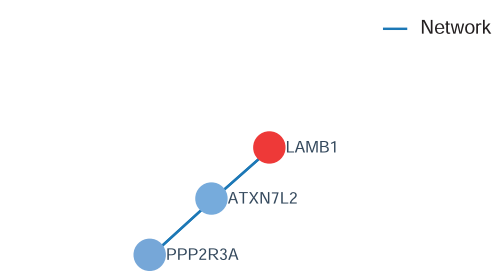

Network
DTX3
RPL19
UBE2L3
TRAF7


1.19
5.22

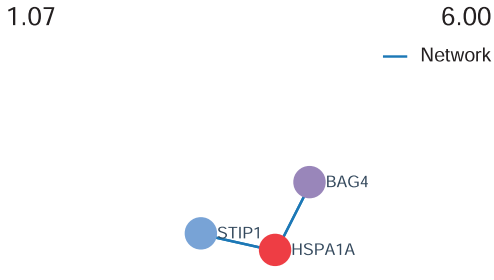

1.25
6.00


Network
PPM1B
USP18
ISG15
UBA7


0.939
6.00


Network
LAMB1
ATXN7L2
PPP2R3A


1.07
6.00


Network
BAG4
STIP1
HSPA1A

0.959
5.10

1.06
6.00

Network
USP4
PRPF3
GTPBP3

1.99
6.00

Network
HLA-DQA1
HLA-DQB1
HCRT

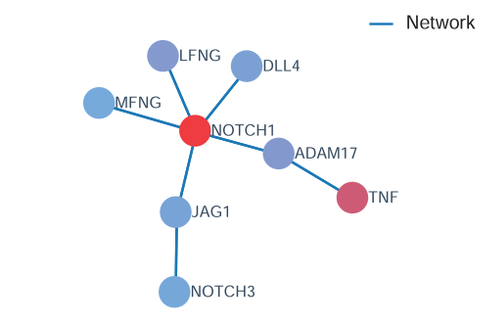
1.18
6.00

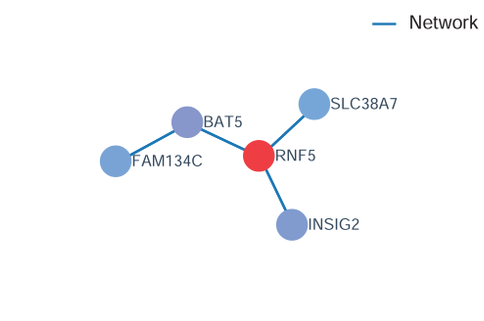
Network
TNPO3
UBE2S
JTB

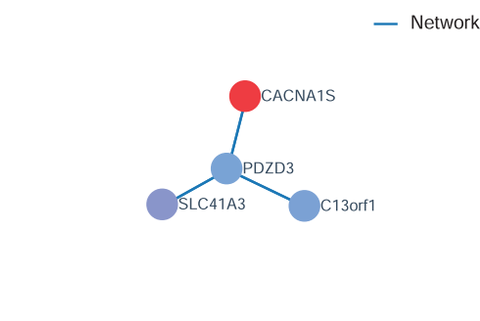
0.971
5.30

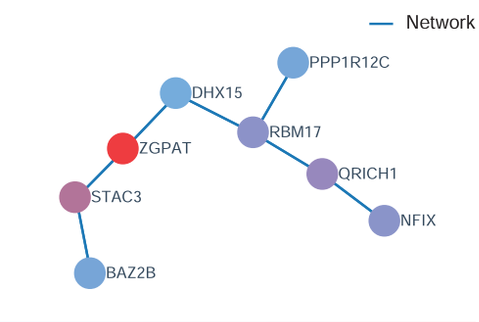
Network
TAP2
TAPBP
TAP1

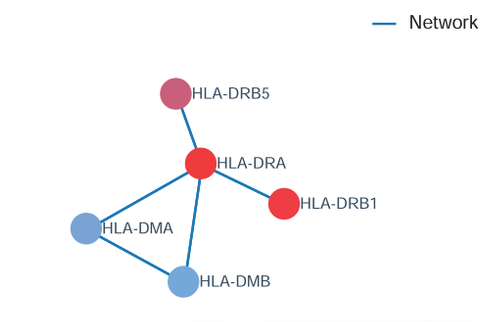
1.06
6.00

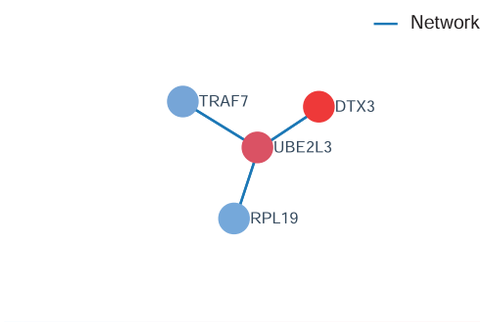

Network
LFNG
DLL4
MFNG
NOTCH1
ADAM17
TNF
JAG1
NOTCH3
1.25
6.66

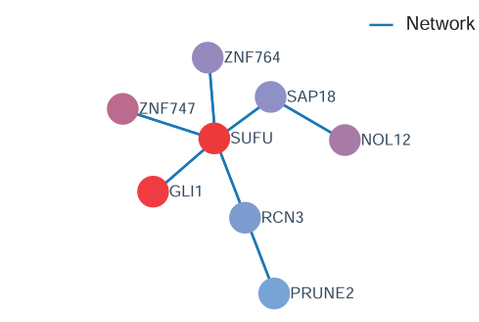

Network
SLC38A7
BAT5
FAM134C
RNF5
INSIG2
0.961
12.0

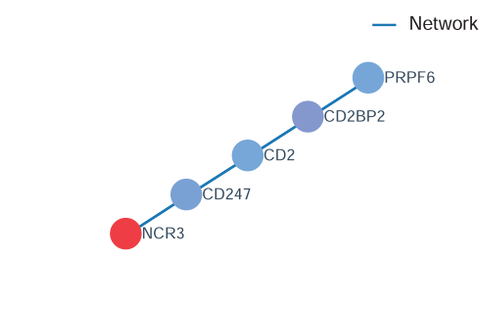

Network
CACNA1S
PDZD3
SLC41A3
C13orf1
0.916
8.40

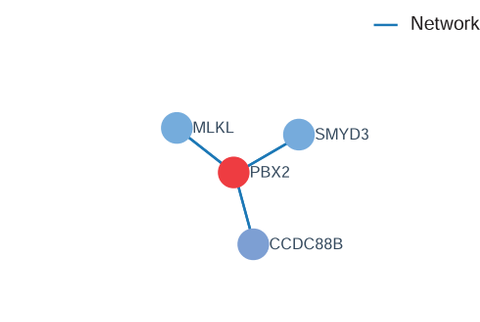

Network
PPP1R12C
DHX15
RBM17
ZGPAT
QRICH1
STAC3
NFIX
BAZ2B
1.13
7.15

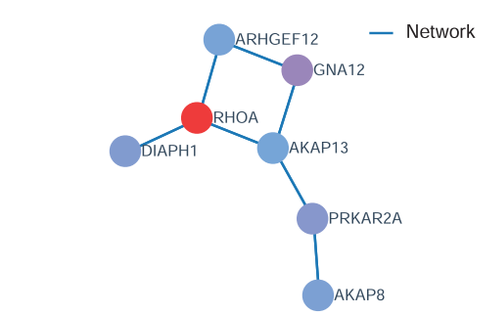

Network
HLA-DRB5
HLA-DRA
HLA-DRB1
HLA-DMA
HLA-DMB
4.60
15.7

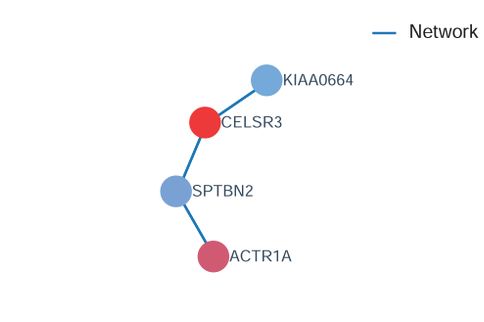

Network
TRAF7
DTX3
UBE2L3
RPL19
1.20
5.47

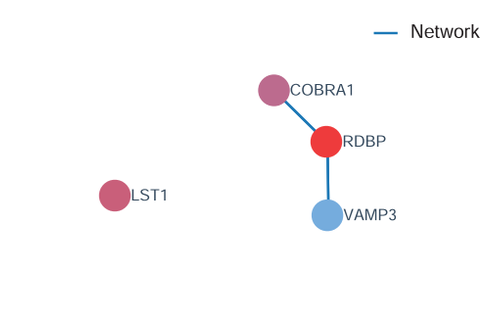

Network
ZNF764
ZNF747
SAP18
SUFU
NOL12
GLI1
RCN3
PRUNE2
0.982
5.95

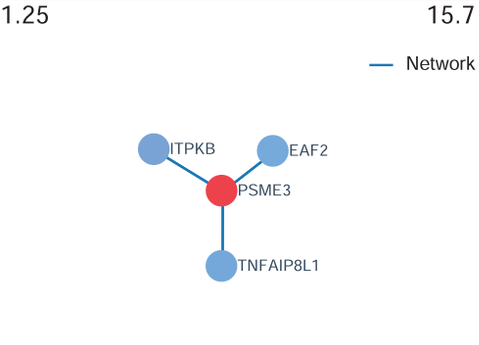

Network
PRPF6
CD2BP2
CD2
CD247
NCR3
1.36
6.51

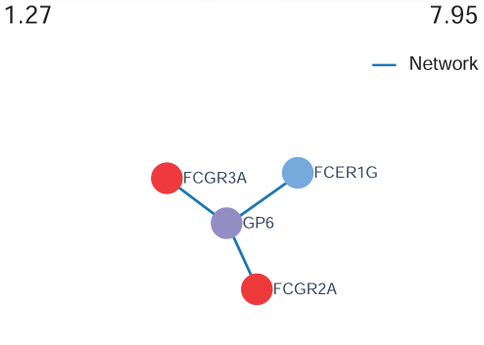

Network
MLKL
SMYD3
PBX2
CCDC88B
1.07
11.4

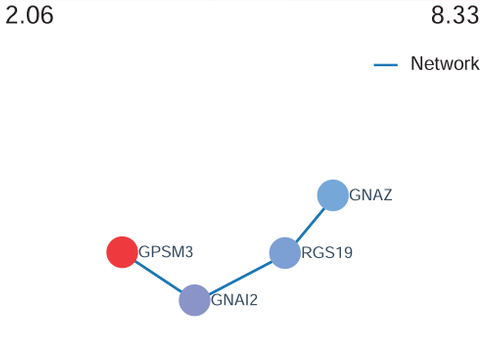

Network
ARHGEF12
GNA12
RHOA
DIAPH1
AKAP13
PRKAR2A
AKAP8
1.25
15.7

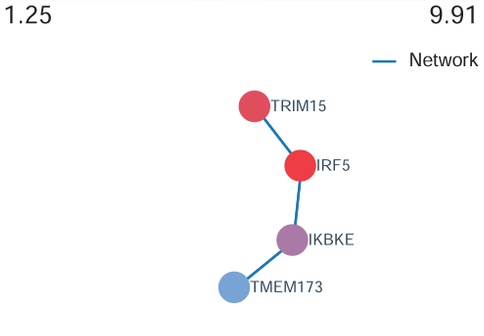

Network
KIAA0664
CELSR3
SPTBN2
ACTR1A
1.27
7.95

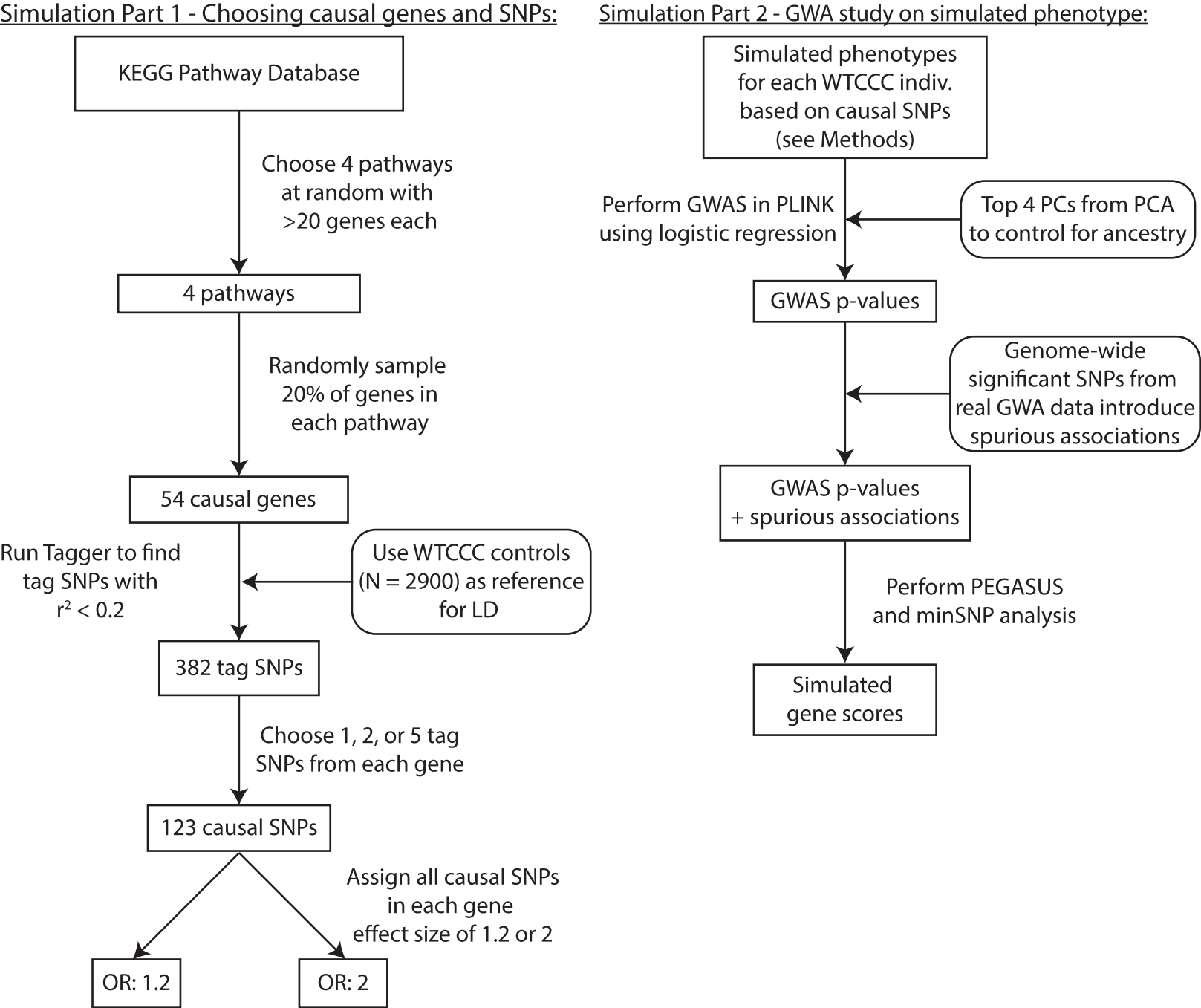

Network
COBRA1
RDBP
LST1
VAMP3
2.06
8.33

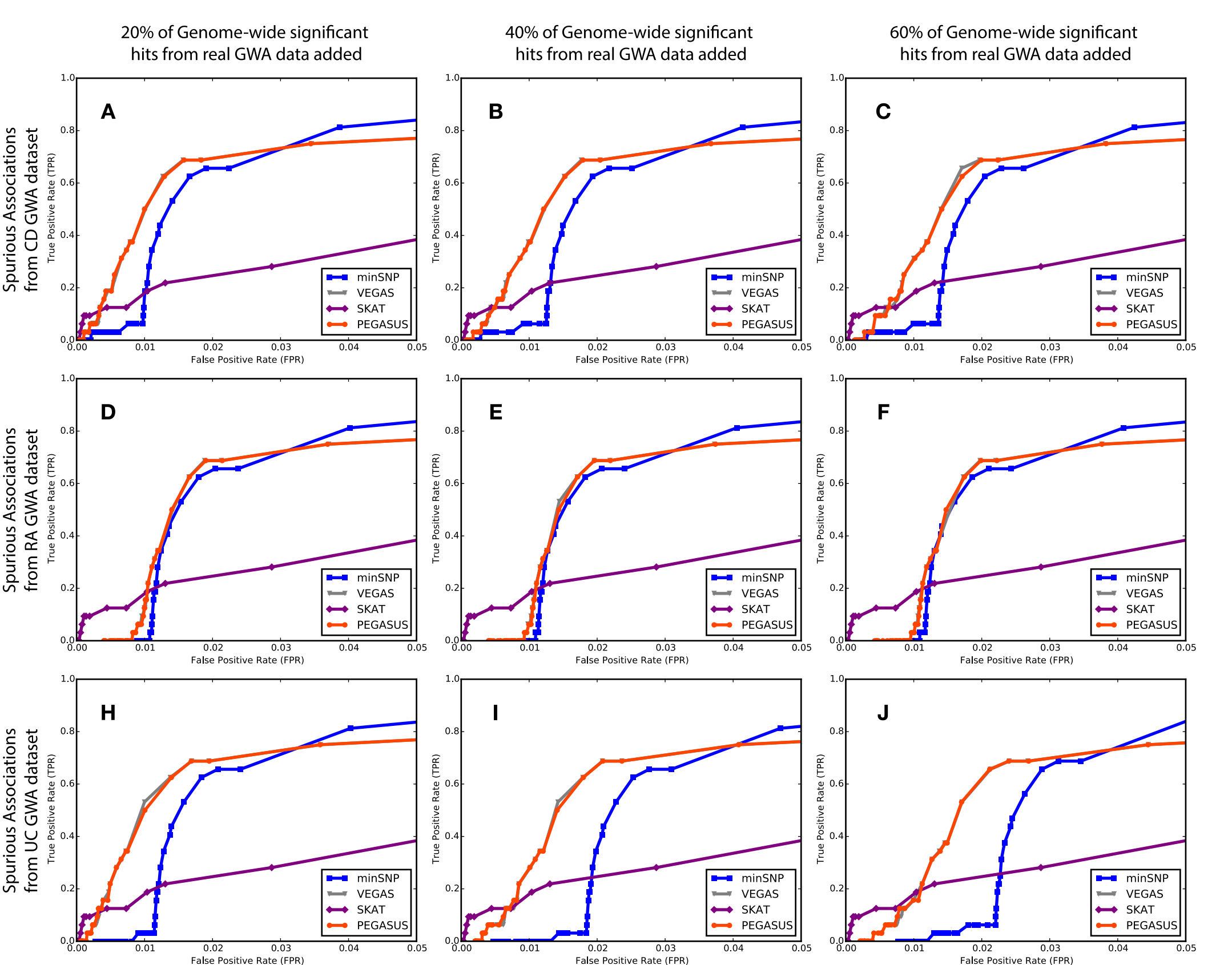

Network
ITPKB
EAF2
PSME3
TNFAIP8L1
1.28
8.05

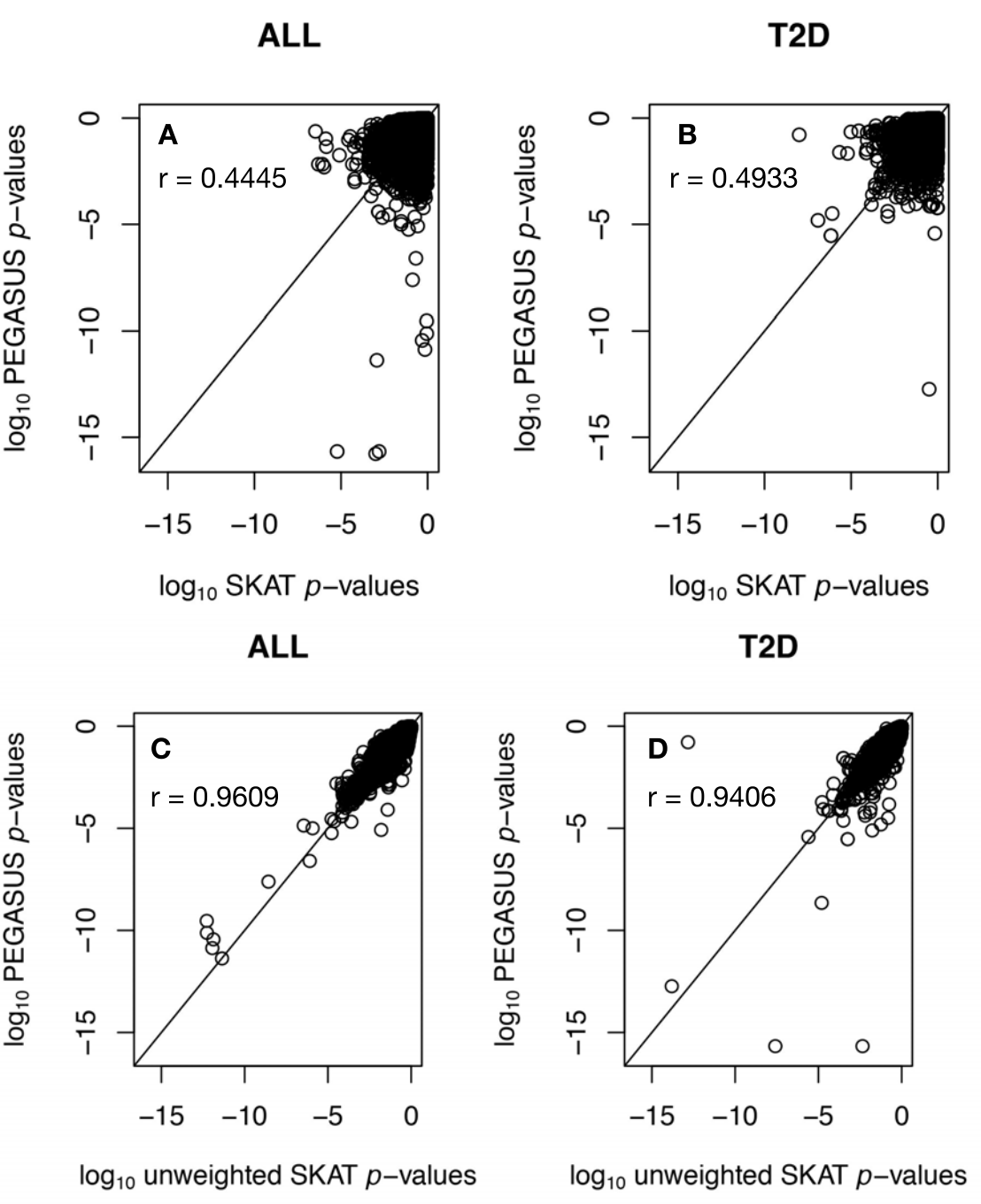

Network
FCGR3A
FCER1G
GP6
FCGR2A
1.25
9.91

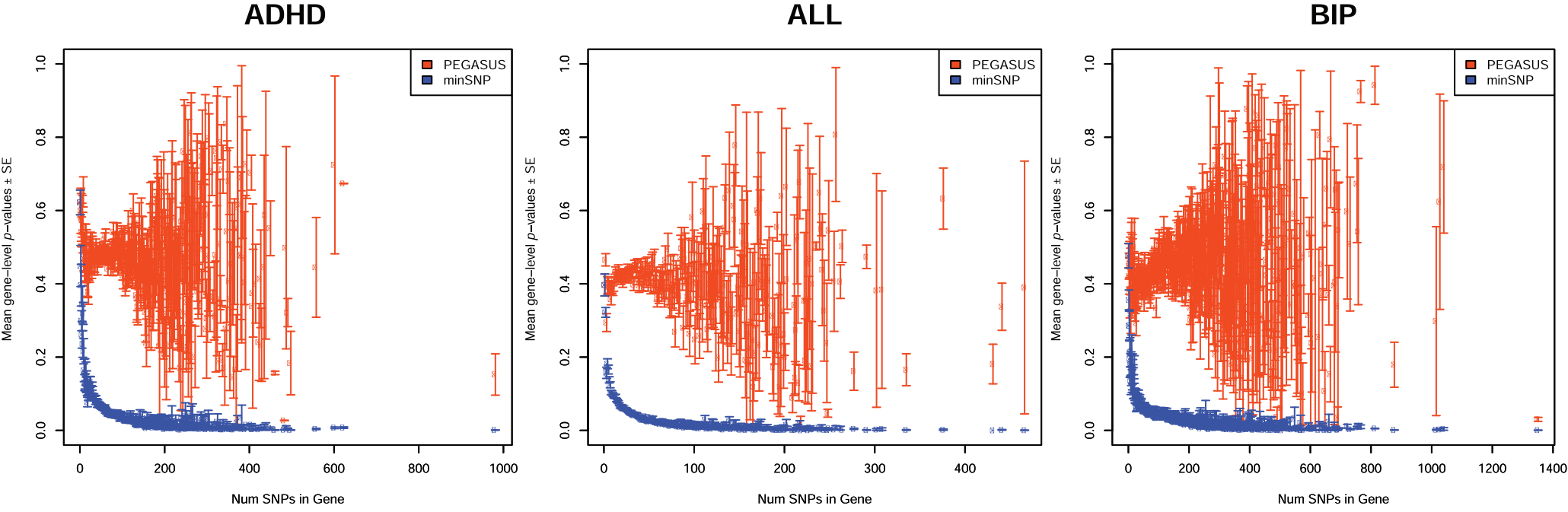

Network
GNAZ
GPSM3
RGS19
GNAI2
1.60
11.5


Network
TRIM15
IRF5
IKBKE
TMEM173
2.45
6.13


Simulation Part 1 - Choosing causal genes and SNPs:
KEGG Pathway Database
Choose 4 pathways at random with >20 genes each
4 pathways
Randomly sample 20% of genes in each pathway
54 causal genes
Run Tagger to find tag SNPs with $r^2 < 0.2$
Use WTCCC controls (N = 2900) as reference for LD
382 tag SNPs
Choose 1, 2, or 5 tag SNPs from each gene
123 causal SNPs
Assign all causal SNPs in each gene effect size of 1.2 or 2
OR: 1.2
OR: 2
Simulation Part 2 - GWA study on simulated phenotype:
Simulated phenotypes for each WTCCC indiv. based on causal SNPs (see Methods)
Perform GWAS in PLINK using logistic regression
Top 4 PCs from PCA to control for ancestry
GWAS p-values
Genome-wide significant SNPs from real GWA data introduce spurious associations
GWAS p-values + spurious associations
Perform PEGASUS and minSNP analysis
Simulated gene scores


20% of Genome-wide significant hits from real GWA data added
40% of Genome-wide significant hits from real GWA data added
60% of Genome-wide significant hits from real GWA data added
Spurious Associations from CD GWA dataset
Spurious Associations from RA GWA dataset
Spurious Associations from UC GWA dataset
A
B
C
D
E
F
H
I
J
True Positive Rate (TPR)
False Positive Rate (FPR)
minSNP
VEGAS
SKAT
PEGASUS


ALL
T2D
A
r = 0.4445
B
r = 0.4933
C
r = 0.9609
D
r = 0.9406
log10 PEGASUS p−values
log10 SKAT p−values
log10 unweighted SKAT p−values
0
−5
−10
−15


ADHD
ALL
BIP
BMI
CD
Height
PEGASUS
minSNP
Mean gene−level p−values ± SE
Num SNPs in Gene

MDD
RA
SCZ
T2D
UC
WHR
PEGASUS
minSNP
Mean gene−level p−values ± SE
Num SNPs in Gene

Table S1

Disease or Trait	Level of Significance for gene score ($\times 10^{-6}$)	# of significant genes (minSNP)	% significant genes (minSNP)	# of significant genes (PEGASUS)	% significant genes (PEGASUS)	# of significant genes (VEGAS)	% significant genes (VEGAS)
ADHD (NEALE *et al.* (2010))	2.81	1	0.01	0	0	0	0
ALL (XU *et al.* (2013))	2.78	29	0.16	10	0.06	10	0.06
BIP (SKLAR *et al.* (2011))	2.81	47	0.26	17	0.1	17	0.1
BMI (SPELIOTES *et al.* (2010))	2.83	96	0.54	56	0.32	56	0.32
CD (FRANKE *et al.* (2010))	2.88	431	2.48	256	1.47	249	1.43
Height (LANGO ALLEN *et al.* (2010))	2.83	834	4.71	404	2.28	397	2.24
MDD (RIPKE *et al.* (2013))	2.81	16	0.09	0	0	0	0
RA (STAHL *et al.* (2010))	2.80	252	1.41	180	1.01	183	1.03
SCZ (RIPKE *et al.* (2011))	2.81	250	1.41	55	0.31	55	0.31
T2D (MORRIS *et al.* (2012))	2.82	54	0.3	11	0.06	10	0.06
UC (ANDERSON *et al.* (2011))	2.80	980	5.5	190	1.07	191	1.07
WHR (HEID *et al.* (2010))	2.83	50	0.28	17	0.1	18	0.1

Number of significant genes and percentage of significant genes (out of the total number of genes considered) for minSNP, PEGASUS and VEGAS gene scores. Significance is based on Bonferroni correction for the total number of genes for which the study had SNP-level p-values in each of 12 GWA datasets (Table 2). We find that minSNP always finds more associated genes than the other two methods. PEGASUS and VEGAS typically find similar numbers of associated genes.

Table S2

Disease or Trait	Reference Data	# WTCCC Cases in Reference Data	# Non-WTCCC Cases in Reference Data	# of PEGASUS Hits in Reference Data replicated in WTCCC data	Total # PEGASUS hits in Reference Data	% PEGASUS hits replicated in WTCCC data (Column 5/Column 6)
BIP	SKLAR *et al.* (2011)	1,571	5,910	0	17	0.0
CD	FRANKE *et al.* (2010)	1,747	4,586	11	254	4.3
RA	STAHL *et al.* (2010)	1,525	4,014	103	180	57.2
T2D	MORRIS *et al.* (2012)	1,924	10,247	1	11	9.1

Replication of PEGASUS gene hits in WTCCC datasets. We conducted a replication experiment for PEGASUS gene hits for the 12 GWA datasets in this study in WTCCC data for Bipolar Disorder (BIP), Crohn's Disease (CD), Rheumatoid Arthritis (RA) and Type 2 Diabetes (T2D). A "hit" is a gene with a gene score $p_g < 2.8 \times 10^{-6}$; this threshold is based on Bonferroni correction for the total number of genes for which the study had SNP-level p-values in each of 12 GWA datasets (Table 2). We find that we are able to replicate up to 57.2% of gene hits in the WTCCC dataset using the PEGASUS method. We note that these four meta-analyses included the WTCCC dataset and are comprised of much larger sample sizes than our replication dataset.

Table S3

Source of GWA p-values	**URL for Downloading p-values**
Psychiatric Genomics Consortium	https://www.med.unc.edu/pgc/downloads
the International IBD Genetics Consortium	http://www.ibdgenetics.org/downloads.html
GIANT consortium	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
Broad Institute	http://www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_etal_2010NG/
DIAGRAM Consortium	http://diagram-consortium.org/downloads.html
Acute Lymphoblastic Leukemia (ALL) data	Genotype data are available from dbGaP. (http://www.ncbi.nlm.nih.gov/gap), and steps to perform the GWA study are outlined in Xu et al (XU *et al.* (2013))

URLs for full SNP-level p-values from GWA datasets analyzed.

SUPPORTING INFORMATION

Text S1

Connection between SKAT and PEGASUS tests

The underlying model for the SKAT test is a multiple linear (or logistic, depending on the phenotype of interest) mixed-model regression, given by Eq. 2; the associated variance component score statistic is given by Eq. 3. Under the null hypothesis of no association, the SKAT model reduces to the following:

$$y_i = \alpha_0 + \boldsymbol{\alpha}\, \boldsymbol{C_i} + \epsilon_i \tag{8}$$

In matrix notation, Eq. 8 can be written in matrix notation as $\mathbf{Y} = \boldsymbol{\alpha}\, \mathbf{C} + \boldsymbol{\epsilon}$, where $\epsilon \quad N(0, \sigma^2)$ and $\mathbf{C}$ is a matrix of covariates, and the estimates of the regression coefficients can be expressed as $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + (\mathbf{C^T C})^{-1}\mathbf{C^T}\boldsymbol{\epsilon}$. We define $\mathbf{P} = \mathbf{C}(\mathbf{C^T C})^{-1}\mathbf{C^T}$, which is a projection matrix as is $\mathbf{I} \quad \mathbf{P}$. We substitute these values into the SKAT test statistic (Eq. 3), which gives the following:

$$\begin{aligned} Q &= [(\boldsymbol{I} \quad \boldsymbol{P})\boldsymbol{\epsilon}]^T \boldsymbol{K}[(\boldsymbol{I} \quad \boldsymbol{P})\boldsymbol{\epsilon}] \\ &= \boldsymbol{\epsilon^T}(\boldsymbol{I} \quad \boldsymbol{P})\boldsymbol{K}(\boldsymbol{I} \quad \boldsymbol{P})\boldsymbol{\epsilon} \\ &= \boldsymbol{\epsilon^T}(\boldsymbol{I} \quad \boldsymbol{P})\boldsymbol{K}\boldsymbol{\epsilon} \end{aligned} \tag{9}$$

Let $d = d_1, \ldots, d_n$ be the eigenvalues of $\sigma^2[(\mathbf{I} \quad \mathbf{P})\mathbf{K}]$. Then, the test statistic takes the form of $\epsilon^{\mathbf{T}} diag(d_1, \ldots, d_n)\epsilon$; this is a quadratic form that follows a mixture of χ^2 distributions with weights given by d.

In PEGASUS, the null distribution is a mixture of χ^2 distributions with weights given by λ, the eigenvalues of the LD matrix Σ. If no covariates are considered and the variant weights are uniform for all variants ($w_j = 1$), the SKAT null distribution becomes a mixture of χ^2 distributions with mixture proportions given by the eigenvalues of the $\mathbf{K} = \mathbf{GWG}$ matrix, which is a variance-covariance matrix similar to the PEGASUS LD matrix $\boldsymbol{\Sigma}$. Thus, under these circumstances, the two tests give similar results (Figure S15C-D).

Text S2

Additional significantly associated gene subnetworks for Attention-Deficit/Hyperactivity Disorder (ADHD), identified using PEGASUS gene scores as input to HotNet2

The subnetwork shown in Figure 5E contains *FURIN*, a gene containing one SNP significantly associated with ADHD (GWA p-value $1.316 \quad 10^{-5}$; NEALE *et al.* (2010)). *FURIN* encodes the protease furin that processes latent precursor proteins into biologically active truncated BDNF molecules in the trans-Golgi network. Decreased levels of truncated BDNF have been shown to be associated with memory loss and learning impairment (CARLINO *et al.* (2013)). Another gene in the subnetwork, *PACS1*, has been found to be moderately associated with years of education in a previous GWA study (p-value $4.9 \quad 10^{-6}$; RIETVELD *et al.* (2013)). PACS1, encoded by *PACS1*, is a trans-golgi-membrane traffic regulator that binds furin; it has been shown to be involved in the localization of the protease furin to the trans-Golgi network (WAN *et al.* (1998)). This gene has also been found to be mutated in cases of an unknown syndrome with intellectual disability (SCHUURS-HOEIJMAKERS *et al.* (2012)). LRP1, or low density lipoprotein receptor-related protein 1, is encoded by the *LRP1* gene and is a major receptor for apolipoprotein E (apoE), which transports lipids to the brain and also plays a role in neuronal repair (FUENTEALBA *et al.* (2010)). LRP1 and apoE are also involved in mediating the clearance of amyloid β 42 from the brain, and defective clearance of amyloid β leads to accumulation in neurons, which contributes to Alzheimer's Disease (FUENTEALBA *et al.* (2010)). Collectively, we find that genes in this subnetwork play important roles in cell trafficking processes that may be involved in the pathology of neuropsychiatric and cognitive disorders such as Alzheimer's Disease and intellectual disability.

A third subnetwork (Figure 5F) contains the genes *RORB* and *MPPED2*, both of which are moderately associated with years of education in previous GWA studies (GWA p-values $7.5 \quad 10^{-4}$ and $6.7 \quad 10^{-4}$, respectively; RIETVELD *et al.* (2013)). *MPPED2* is also associated with information processing speed (LUCIANO *et al.* (2011): GWA p-value $1.6 \quad 10^{-5}$). *MPPED2* is expressed in the fetal brain and is associated with WAGR syndrome whose symptoms include Wilms' tumor, aniridia, genitourinary anomalies and retardation, but the gene's function in the brain is still unclear (CHEN *et al.* (2010); DAVIS *et al.* (2008)). The gene *NR2F2*, which encodes the COUP-TFII transcription factor, is important in many aspects of neural development including neurogene-

sis, axogenesis, and differentiation (NAKA *et al.* (2008)). *RORB* encodes the transcription factor RORB which is known to play a regulatory role in neurogenesis and has been found to be associated with bipolar disorder (MCGRATH *et al.* (2009)). The gene *RARG* encodes a retinoic acid receptor (retinoic acid receptor gamma) which makes a heterodimeric pair with a retinoid X receptor. The RXRG gene encodes one such retinoid X receptor (MADEN (2007)). The retinoic acid receptor-retinoid X receptor heterodimeric pair are involved in mediating response to retinoic acid by binding to a DNA region containing a retinoic acid response element (RARE), which is bound to a corepressor protein and regulates transcription. Retinoic acid is very important in development of the nervous system, especially in patterning and neuronal differentiation (MADEN (2007)). The genes in this subnetwork (Figure 5F) are thus likely involved in brain development. The genes in Figure 5D-F were found using VEGAS gene scores as input to HotNet2 as well.

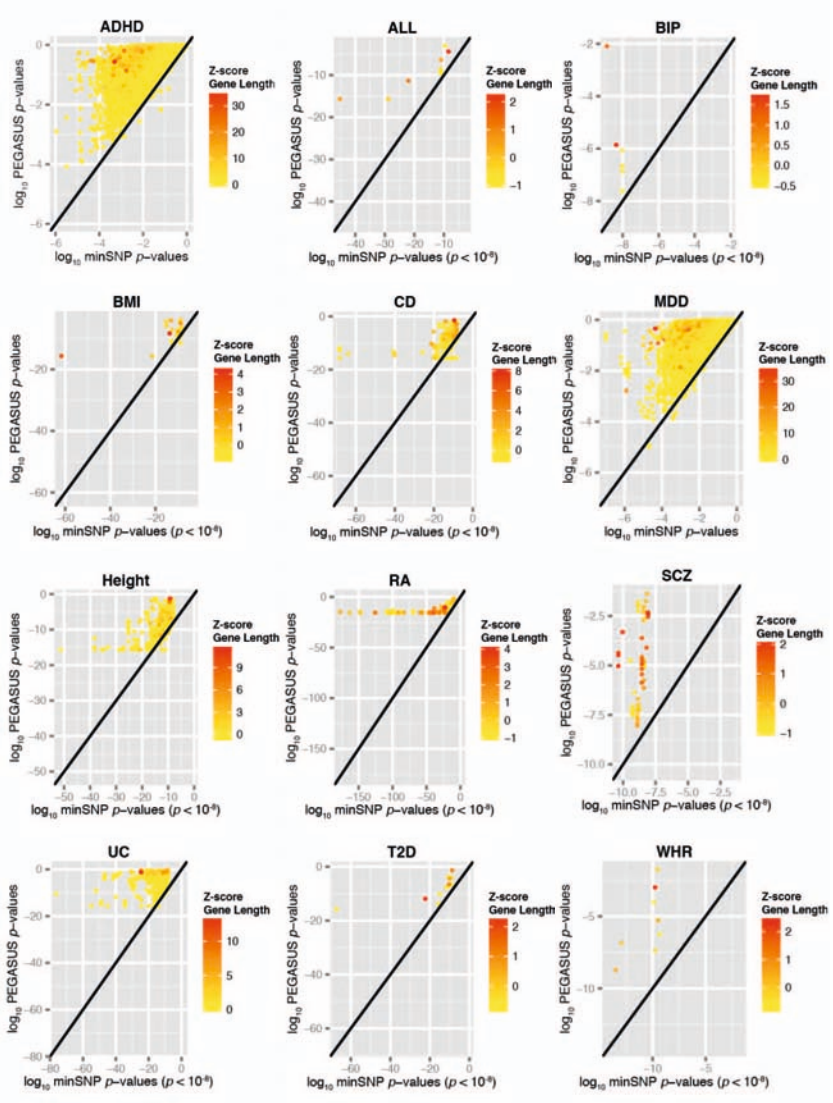
Text S3

Additional significantly associated gene subnetworks for Waist-Hip Ratio (WHR), identified using PEGASUS gene scores as input to HotNet2

The second subnetwork found by HotNet2 (LEISERSON *et al.* (2015)) using PEGASUS gene scores as input contains the gene *VEGFA* (vascular endothelial growth factor A), which is thought to mediate adipogenesis (HEID *et al.* (2010)). Serum levels of VEGFA are correlated with obesity (HEID *et al.* (2010)). VEGFA is also associated with angiogenesis (formation of new blood vessels) and for these reasons, VEGFA has been posited to underlie the association between childhood obesity and increased risk for atherosclerosis (SIERVO *et al.* (2012)). The growth factor encoded by *VEGFB* (vascular endothelial growth factor B), which is also contained in the subnetwork, is involved in endothelial targeting of fatty acids to peripheral tissues. Mice deficient in *VEGFB* show decreased uptake of lipids by heart, muscle and brown adipose tissue, which are mitochondria rich and use fatty acids as an energy source, and increased accumulation of lipids in white adipose tissue instead, ultimately resulting in increased body weight (HAGBERG *et al.* (2010)). *FLT1*, which encodes the vascular endothelial growth factor receptor 1, and NRP1 or neurophilin 1, encoded by the *NRP1* gene, are a receptor and co-receptor, respectively for VEGFB. Mice lacking the *FLT1* gene and mice deficient in *NRP1* both show down-regulation of

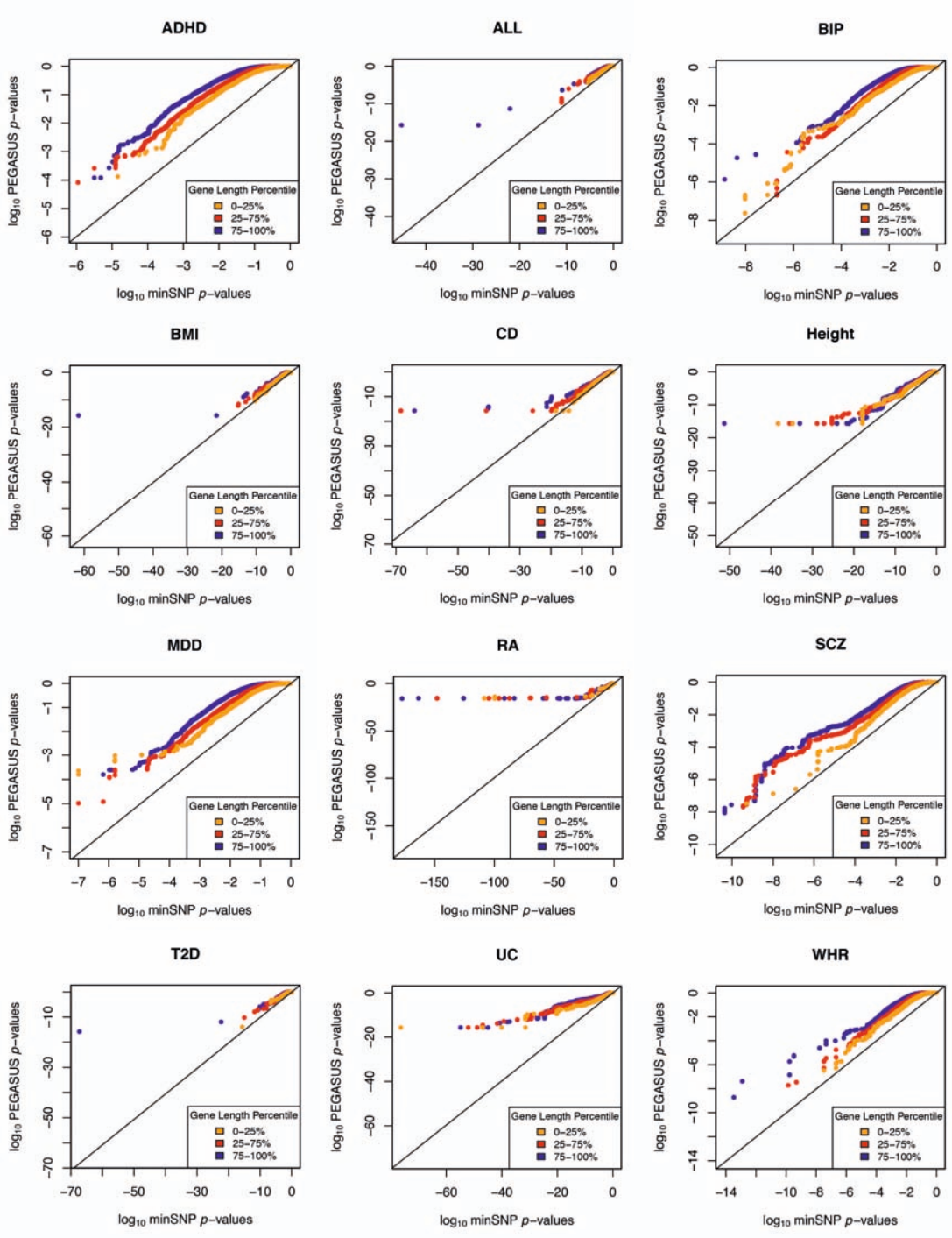
fatty acid transport proteins in the heart, suggesting that VEGFB functions via FLT1 and NRP1 signaling (HAGBERG *et al.* (2010)). In this subnetwork, we see genes known to be associated with obesity and lipid trafficking; taken together, this result elucidates the mechanisms of action of VEGFA, which contains a known GWA study association for WHR (SNP p-value $1.38 \quad 10^{-10}$; HEID *et al.* (2010)) and VEGFB, which is moderately associated with WHR (SNP p-value $7.2 \quad 10^{-3}$; HEID *et al.* (2010)). These genes were found using VEGAS gene scores as input to HotNet2 as well.

Figure S1



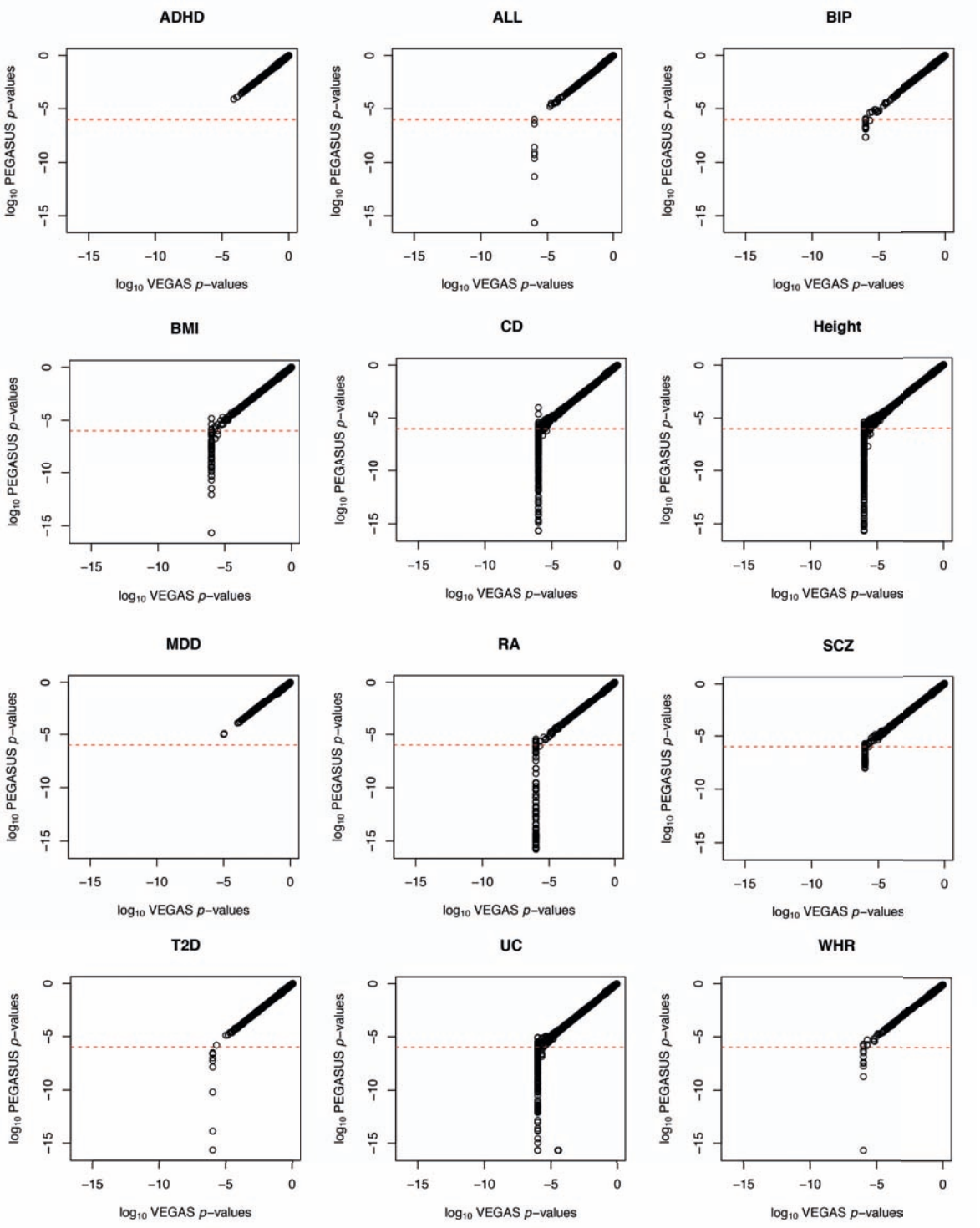
Log-log plots of minSNP gene scores ($< 10^{-8}$ where possible) and corresponding PEGASUS gene scores for the 12 datasets analyzed in this study (Table 2). The points are colored by the z-score for the number of SNPs in the gene in each study, where red represents the highest possible z-score and yellow the lowest. We find that minSNP gene scores are almost always lower then PEGASUS gene scores.

Figure S2



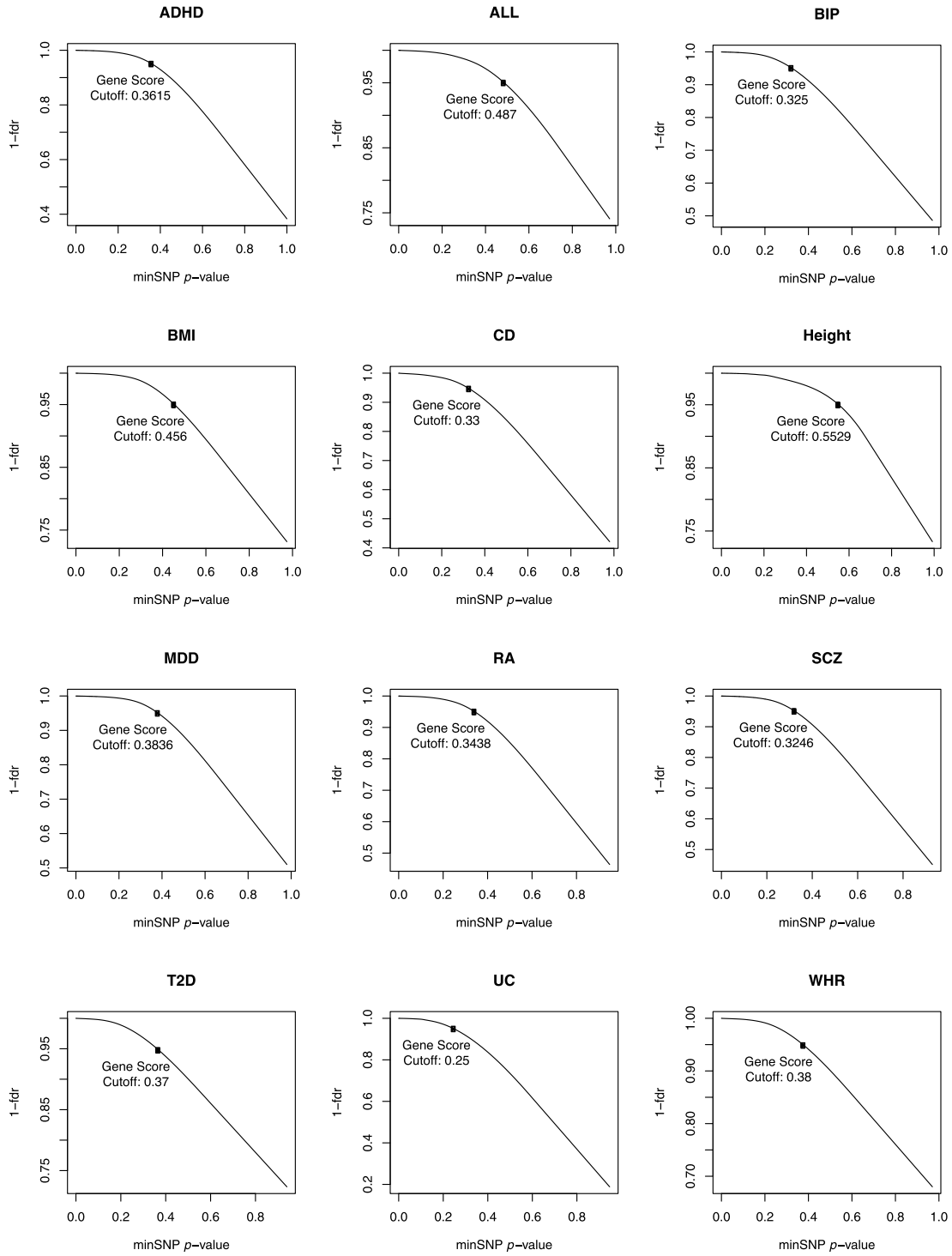
Quantile-quantile plots comparing the base-10 logarithm of PEGASUS gene scores against the base-10 logarithm of minSNP gene scores for the 12 datasets analyzed in this study (Table 2). Each point represents a gene and is colored yellow, red or blue based on gene length (measured by the number of SNPs in the gene $\pm$ 50 kb boundary) percentile: 0-25%, 25-75%, and 75-100%, respectively. We find that minSNP gene scores are smaller than PEGASUS gene scores and decrease with increasing number of SNPs in a gene.

Figure S3



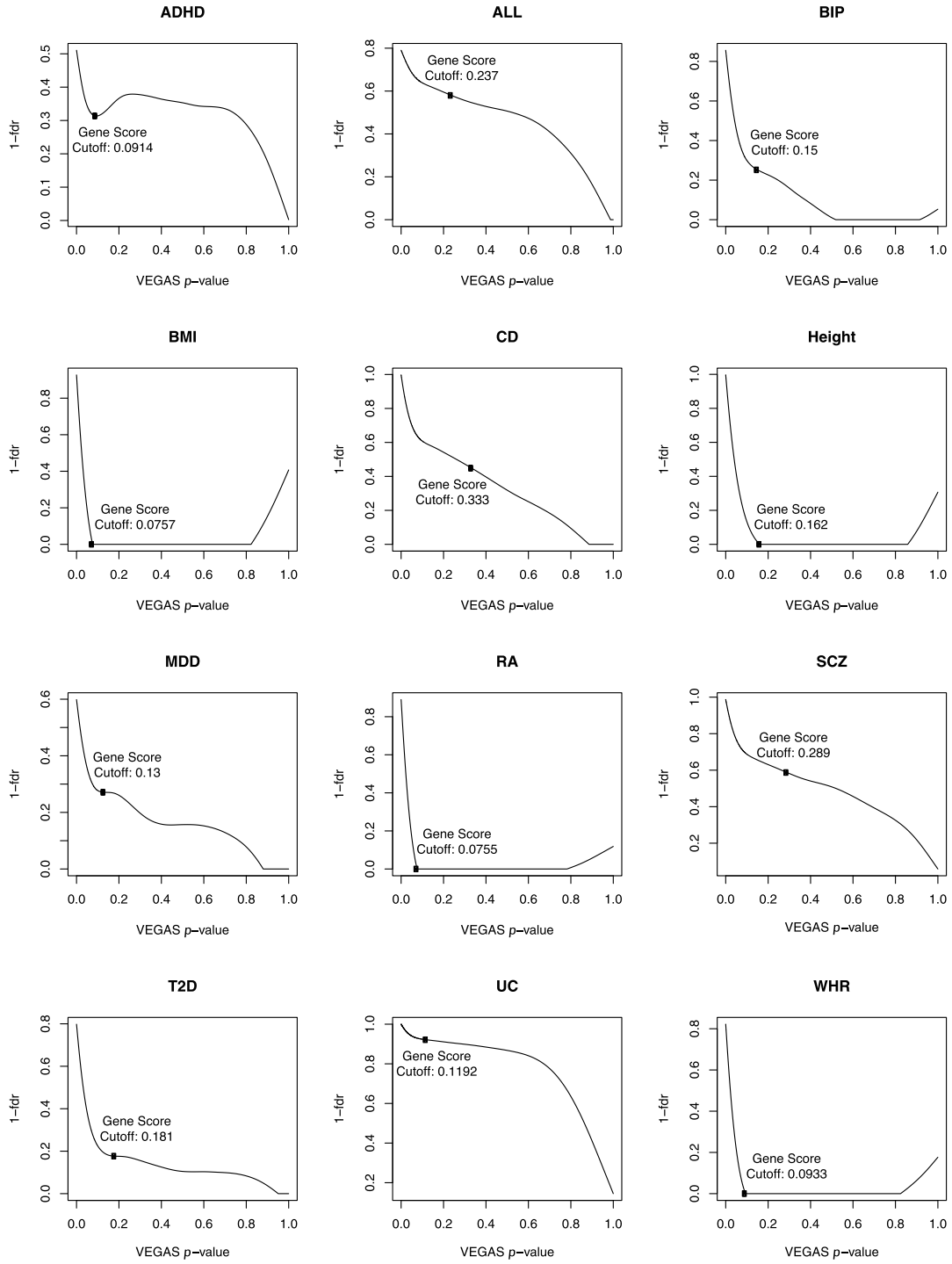
Log-log (base-10 logarithm) plots of PEGASUS against VEGAS gene scores for the 12 datasets analyzed in this study (Table 2). There is high concordance between gene score results from the two methods for all genes except those with gene scores $\leq 10^{-6}$ (below red dashed line); such genes can be ranked by gene score using PEGASUS, but not VEGAS.

Figure S4



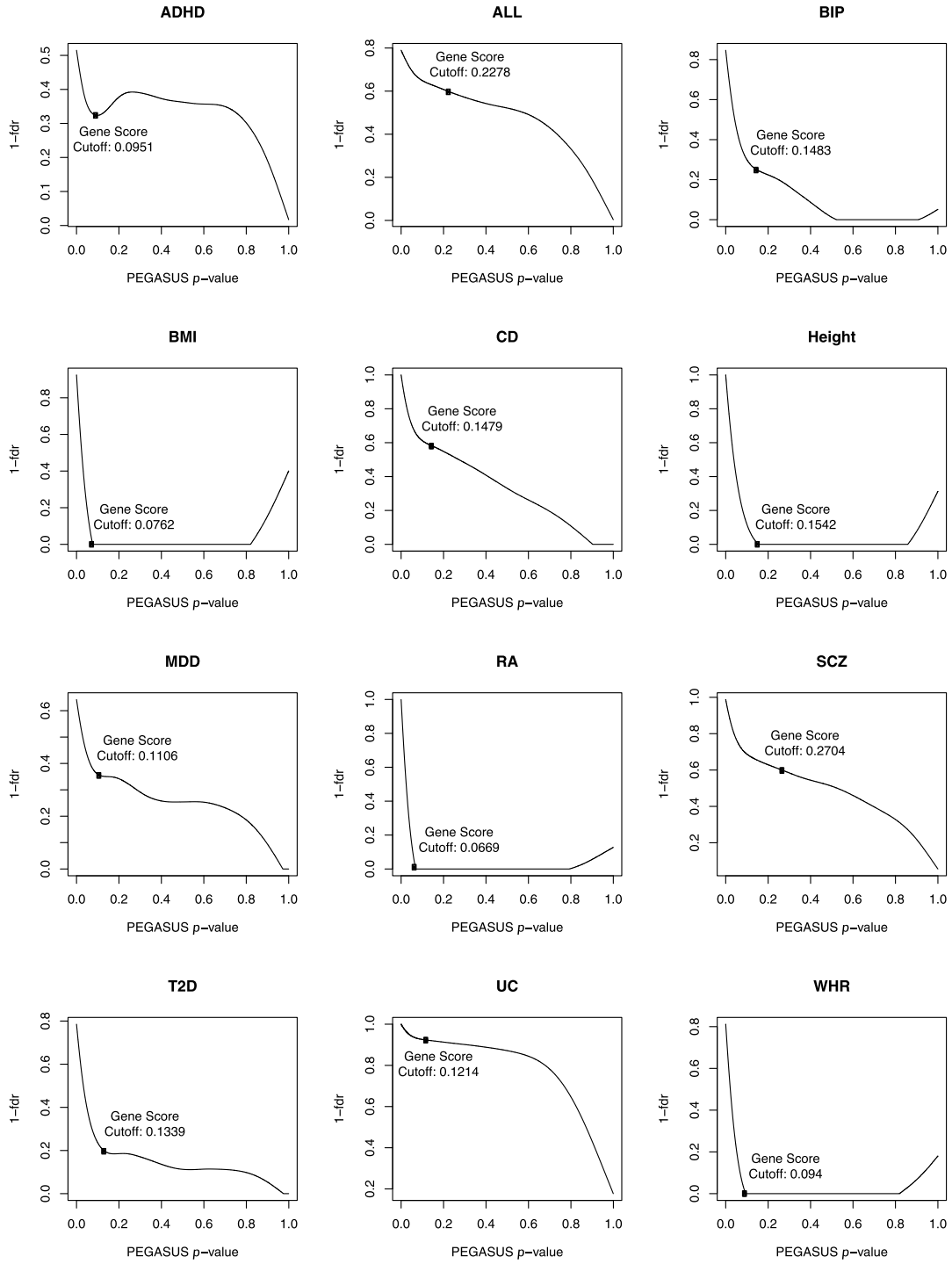
1 - Local FDR (lFDR) versus minSNP gene scores with gene score threshold indicated at lFDR of 0.05 for the 12 datasets analyzed in this study (Table 2). The black squares represent the gene score threshold for minSNP gene scores used as input to HotNet2 analysis. The gene score threshold is determined by finding the gene score where lFDR reaches 0.05.

Figure S5



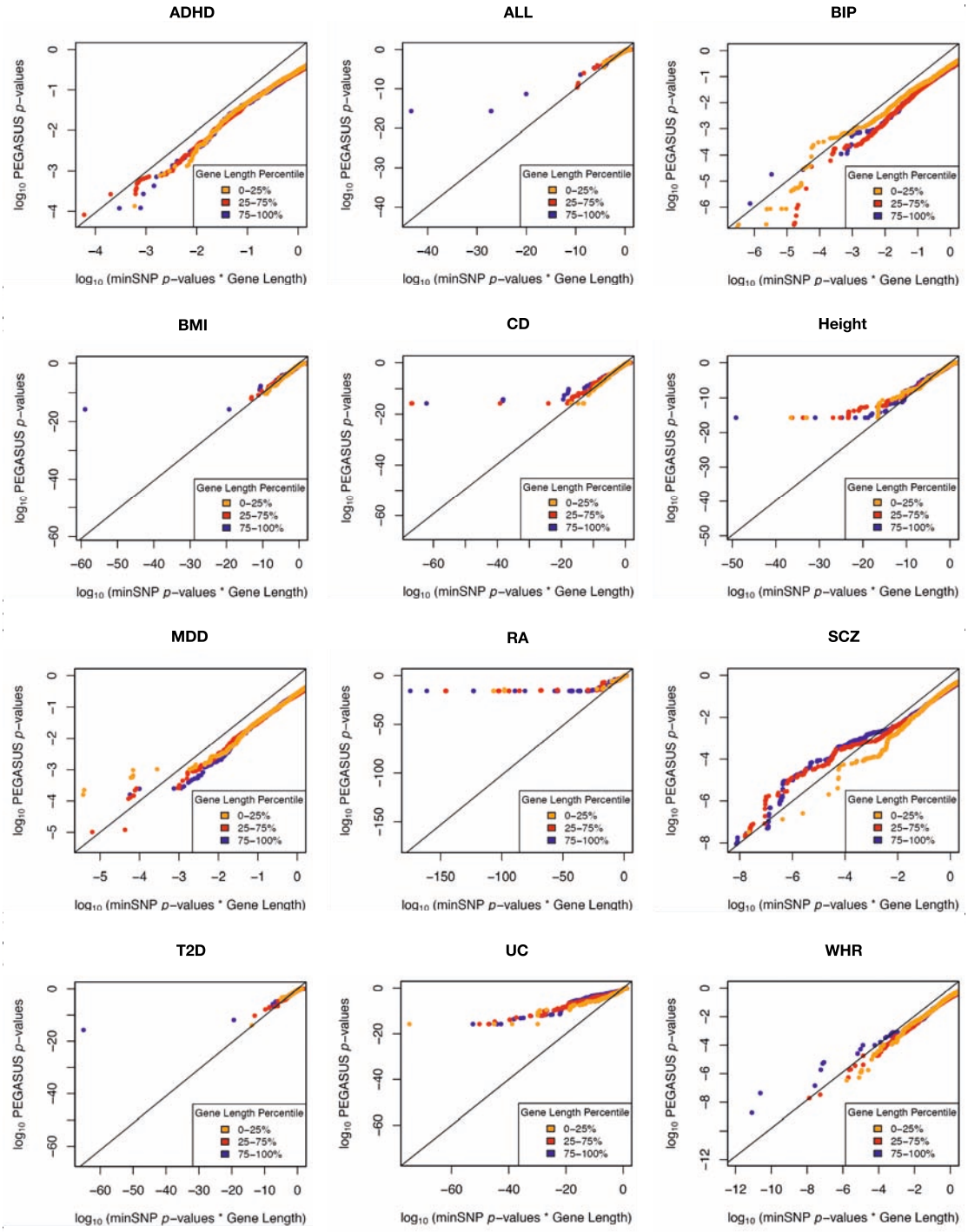
1 - Local FDR (IFDR) versus VEGAS gene scores with gene score threshold indicated at the first "elbow" point in the graph for the 12 datasets analyzed in this study (Table 2). The black squares represent the gene score threshold for VEGAS gene scores used as input to HotNet2 analysis. The gene score threshold is determined by finding the first "elbow" point in the graph.

Figure S6



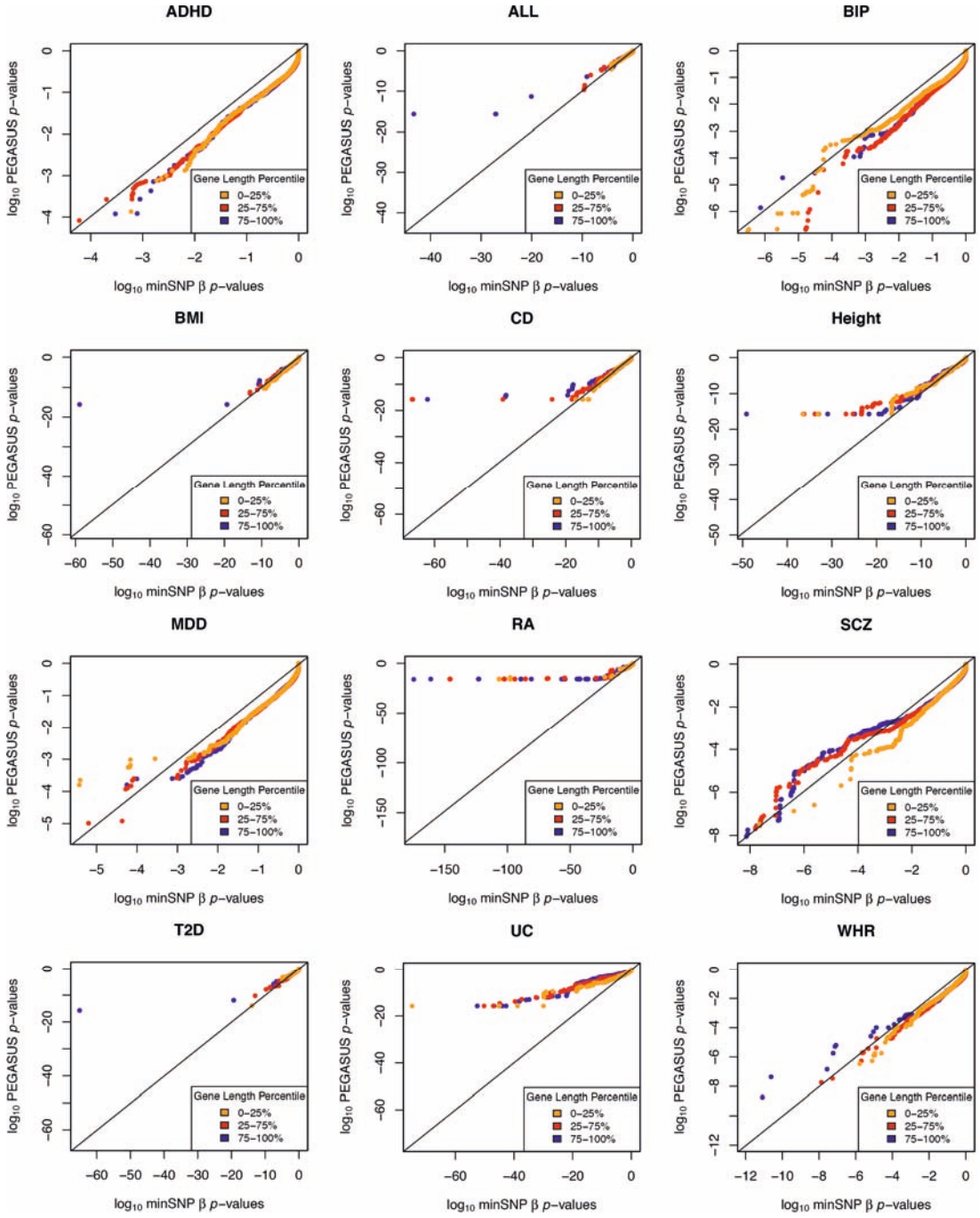
1 - Local FDR (lFDR) versus PEGASUS gene scores with gene score threshold indicated at the first "elbow" point in the graph for the 12 datasets analyzed in this study (Table 2). The black squares represent the gene score threshold for PEGASUS gene scores used as input to HotNet2 analysis. The gene score threshold is determined by finding the first "elbow" point in the graph.

Figure S7



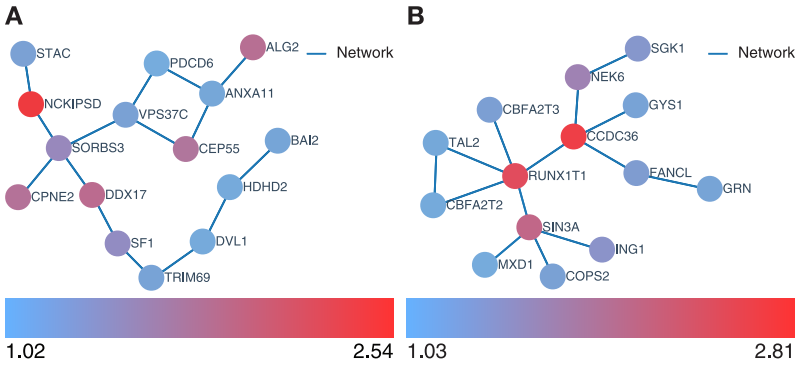
Quantile-quantile plots comparing the base-10 logarithm of PEGASUS gene scores against the base-10 logarithm of minSNP gene scores, multiplied by gene length. Each point represents a gene and is colored yellow, red or blue based on gene length (measured by the number of SNPs in the gene $\pm$ 50 kb boundary) percentile: 0-25%, 25-75%, and 75-100%, respectively. minSNP gene scores, when corrected for gene length by multiplying by the number of SNPs in a gene, are still lower for genes with higher gene length.

Figure S8



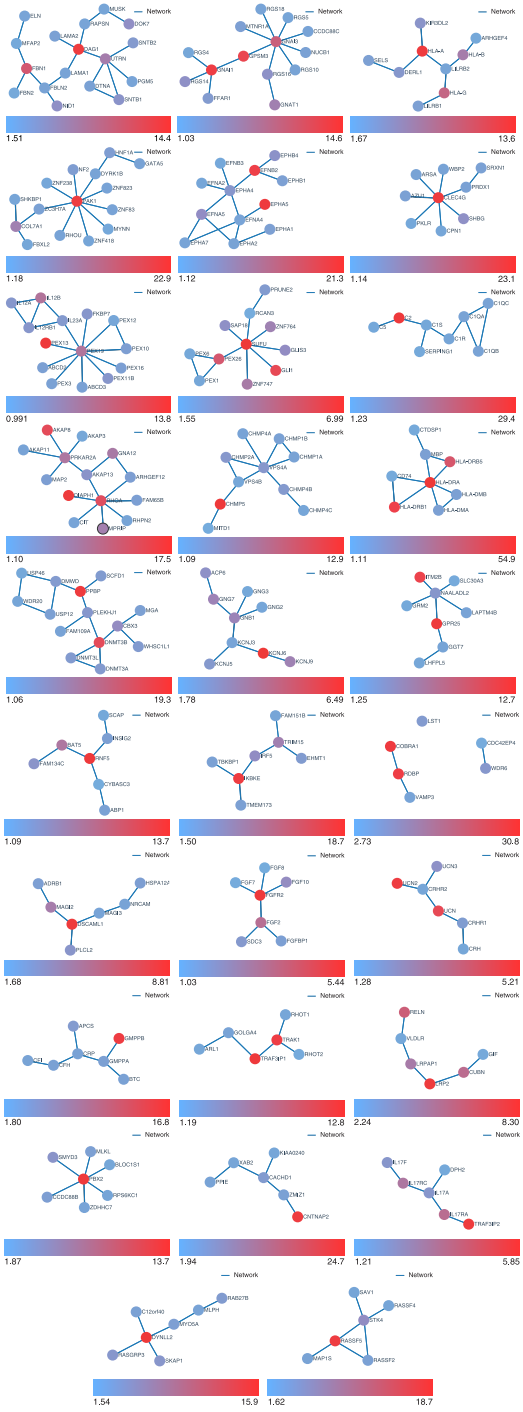
Quantile-quantile plots comparing the base-10 logarithm of PEGASUS gene scores against the base-10 logarithm of minSNP gene scores from the Beta(1, gene length) distribution for the 12 datasets analyzed in this study (Table 2). Each point represents a gene and is colored yellow, red or blue based on gene length (measured by the number of SNPs in the gene $\pm$ 50 kb boundary) percentile: 0-25%, 25-75%, and 75-100%, respectively. minSNP gene scores, when corrected for gene length by calculating a p-value for the gene score from the Beta(1, gene length) distribution, are still lower for genes with higher gene length.

Figure S9



Additional significantly associated gene subnetworks for ADHD identified using PEGASUS gene scores and HotNet2. Circles represent genes in each subnetwork and are colored by heat score (negative log-transformed PEGASUS gene scores); the color bar indicates the lowest heat score (blue or "cold" genes) and highest heat score (red or "hot" genes) in each subnetwork. Lines between genes indicate a gene interaction from the HINT database (DAS and YU (2012)).

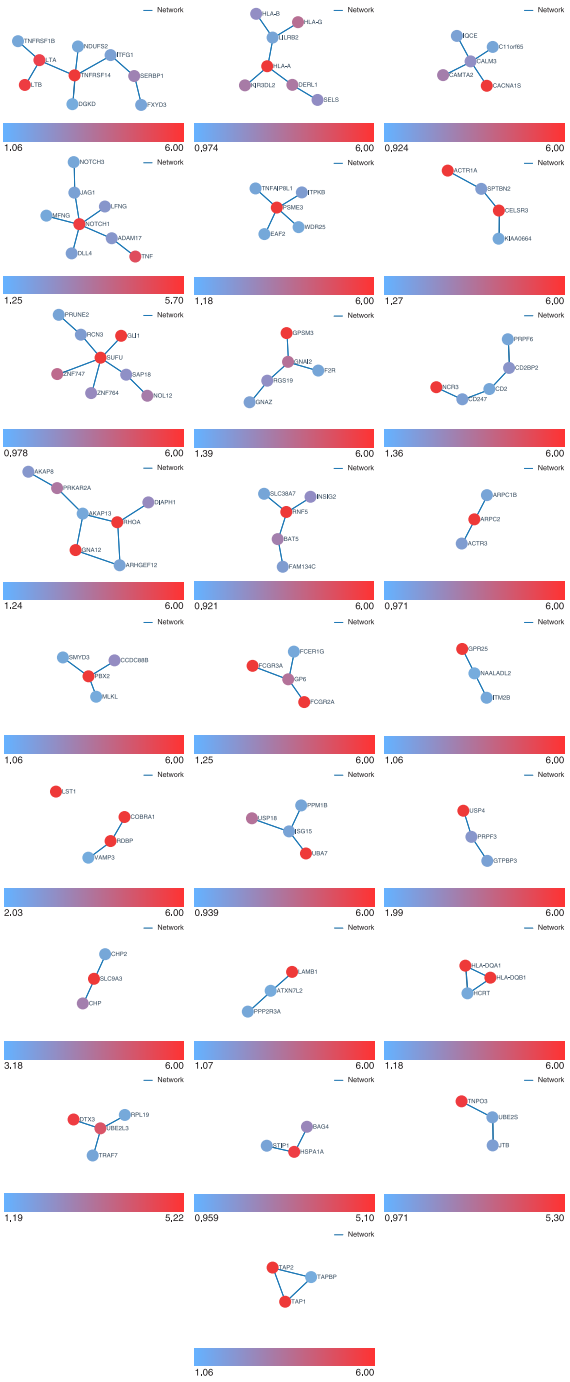
Figure S10



HotNet2 results for Ulcerative Colitis (UC) using minSNP gene scores. We performed HotNet2 analysis using minSNP gene scores for Ulcerative Colitis and found these significant gene interaction subnetworks ($p \leq 0.05$). Subnetworks containing only genes with p_g 10^{-5} are not shown in this figure. Circles represent genes in each subnetwork and are colored by heat score (negative log-transformed minSNP gene scores); the color bar indicates the lowest heat score (blue or "cold" genes) and highest heat score (red or "hot" genes) in each subnetwork. Lines between genes indicate a gene interaction from the HINT database (DAS and YU (2012)). minSNP

subnetworks typically show “star-shaped” networks with many low scoring genes that are likely artifacts.

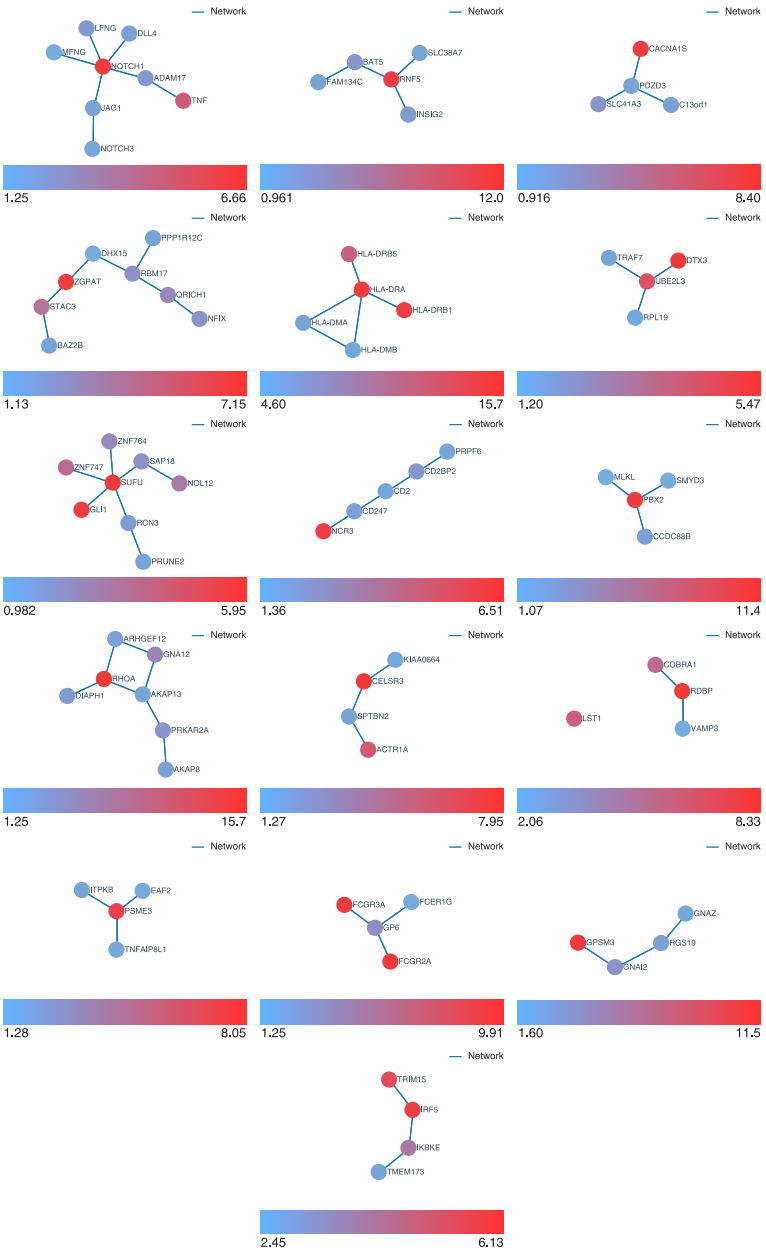
Figure S11



HotNet2 results for Ulcerative Colitis (UC) using VEGAS gene scores. We performed HotNet2 analysis using VEGAS gene scores for Ulcerative Colitis and found these significant gene interaction subnetworks ($p \leq 0.05$). Subnetworks containing only genes with $p_g \quad 10^{-5}$ are not shown in this figure. Circles represent genes in each subnetwork and are colored by heat score (negative

log-transformed VEGAS gene scores); the color bar indicates the lowest heat score (blue or "cold" genes) and highest heat score (red or "hot" genes) in each subnetwork. Lines between genes indicate a gene interaction from the HINT database (DAS and YU (2012)). These subnetworks are missing potentially biologically relevant results that can be found with PEGASUS due to its higher precision gene scores (Figure 4).

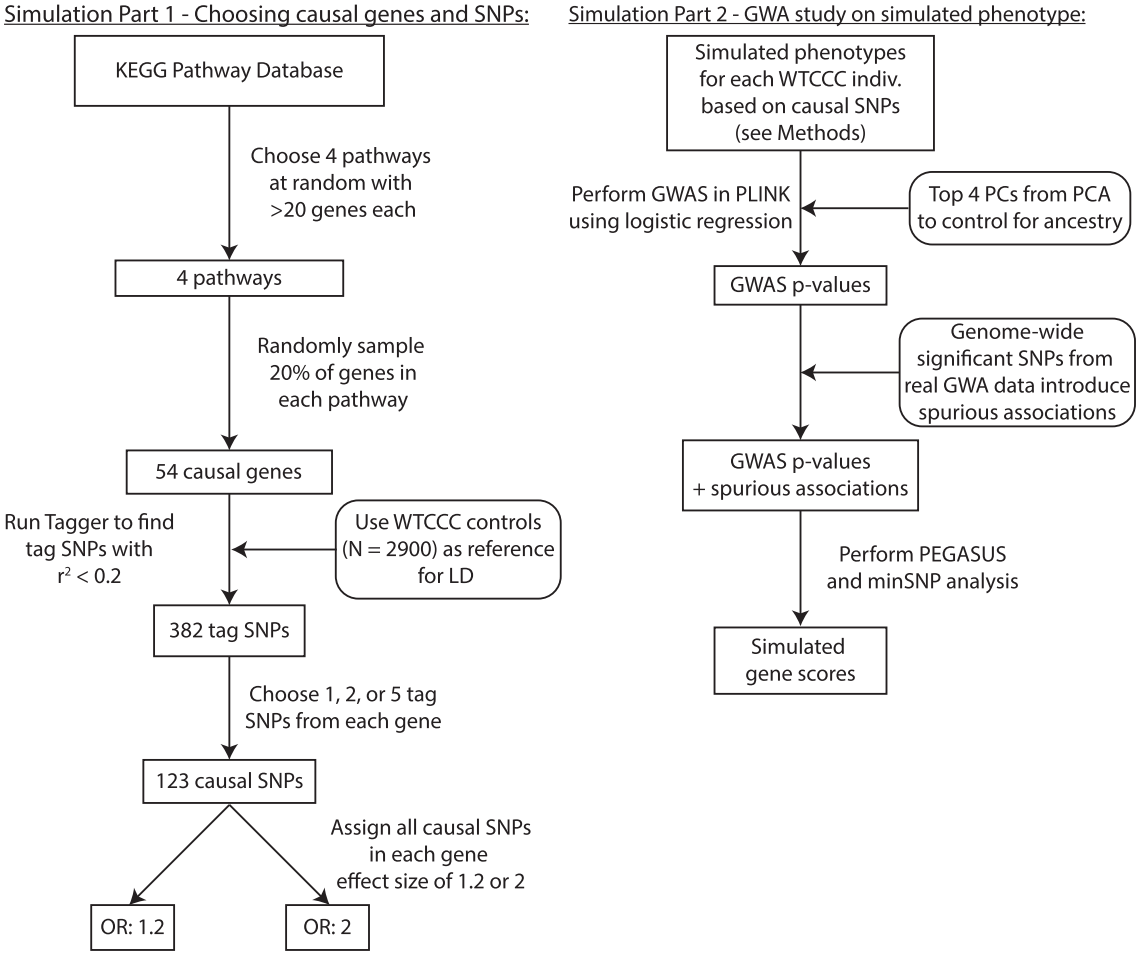
Figure S12



Additional significantly associated subnetworks for Ulcerative Colitis (UC) identified using PEGASUS gene scores and HotNet2. We performed HotNet2 analysis using PEGASUS gene scores for Ulcerative Colitis and found these significant gene interaction subnetworks ($p \leq 0.05$) in addition to the subnetworks in Figure 5A-C. Subnetworks containing only genes with $p_g \quad 10^{-5}$ are not shown in this figure. Circles represent genes in each subnetwork and are colored by heat

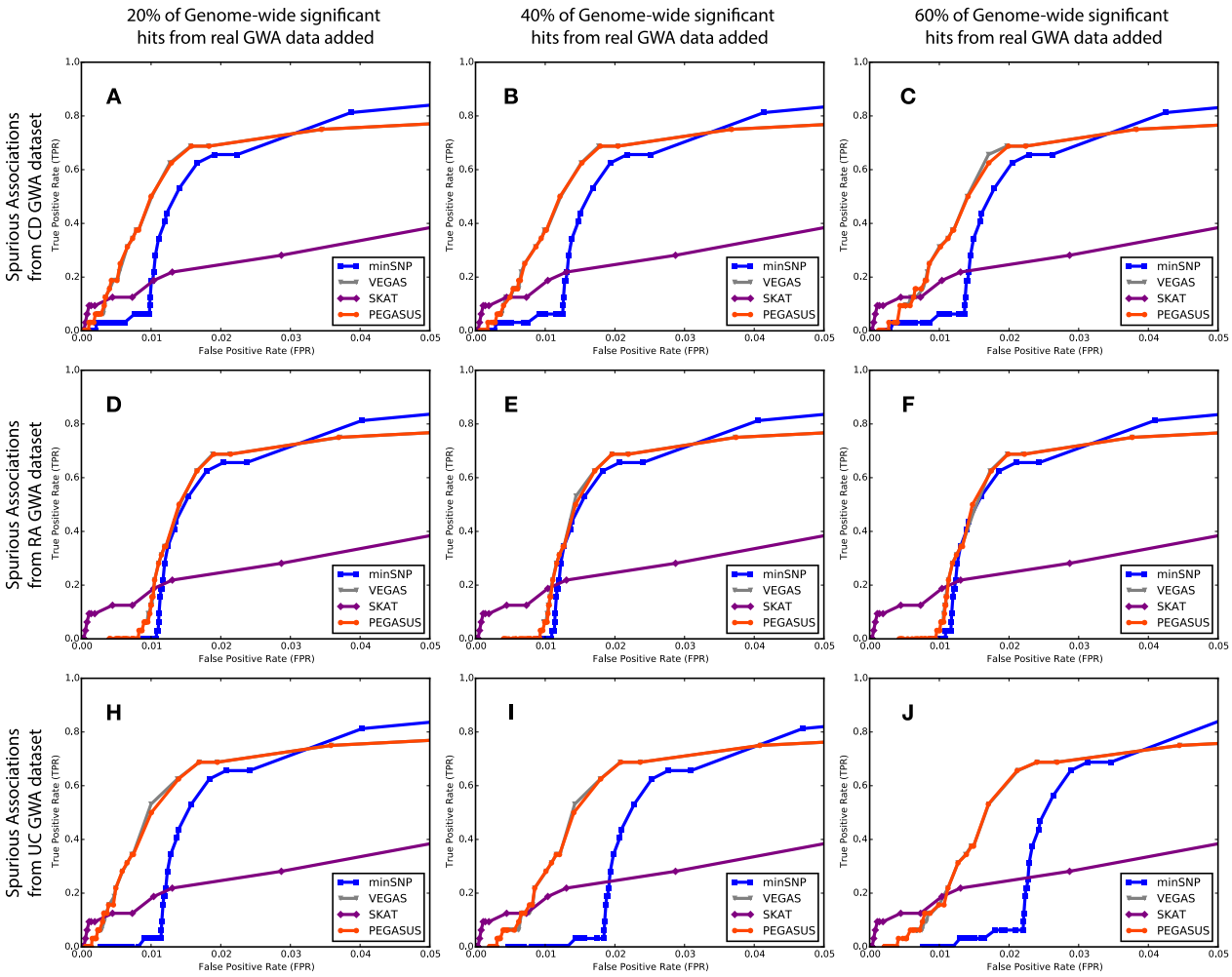
score (negative log-transformed PEGASUS gene scores); the color bar indicates the lowest heat score (blue or "cold" genes) and highest heat score (red or "hot" genes) in each subnetwork. Lines between genes indicate a gene interaction from the HINT database (DAS and YU (2012)).

Figure S13



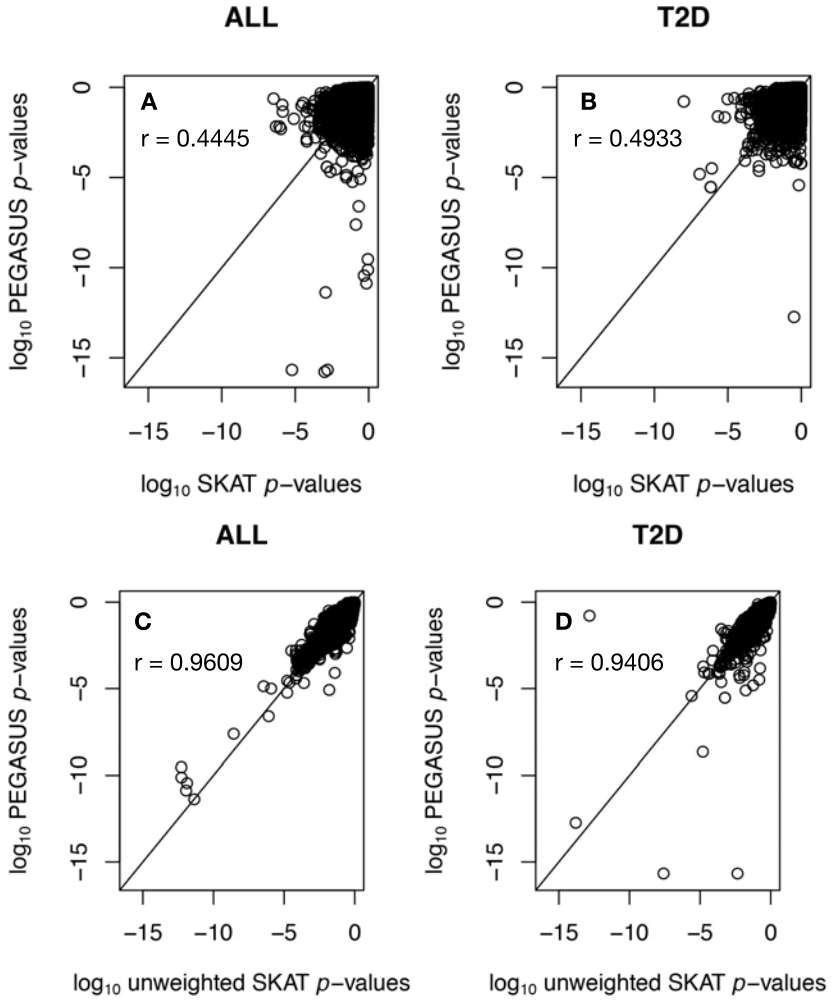
Steps to simulate GWA data with known genomic architecture. We performed a GWA study for simulated phenotype data with known underlying true causal genes based on the approach described in Wojcik et al (WOJCIK *et al.* (2015)). We first choose casual genes from the KEGG Pathway Database (KANEHISA (1997); KANEHISA *et al.* (2012)), find tag SNPs using WTCCC control individuals as reference for LD (WTCCC (2007)), choose 1,2,or 5 tag SNPs from each gene to be causal SNPs and then randomly assign all SNPs in a gene an odds ratio of 1.2 or 2. We assign case/control status to the WTCCC control individuals based on their minor allele dosage at each causal SNP (see Subjects and Methods), and we perform logistic regression on minor allele dosage and simulated phenotype, using the top four principal components as covariates in the logistic regression to control for ancestry, to obtain GWA SNP-level p-values. Simulated gene-level p-values are then generated from SNP-level using PEGASUS, VEGAS and minSNP.

Figure S14



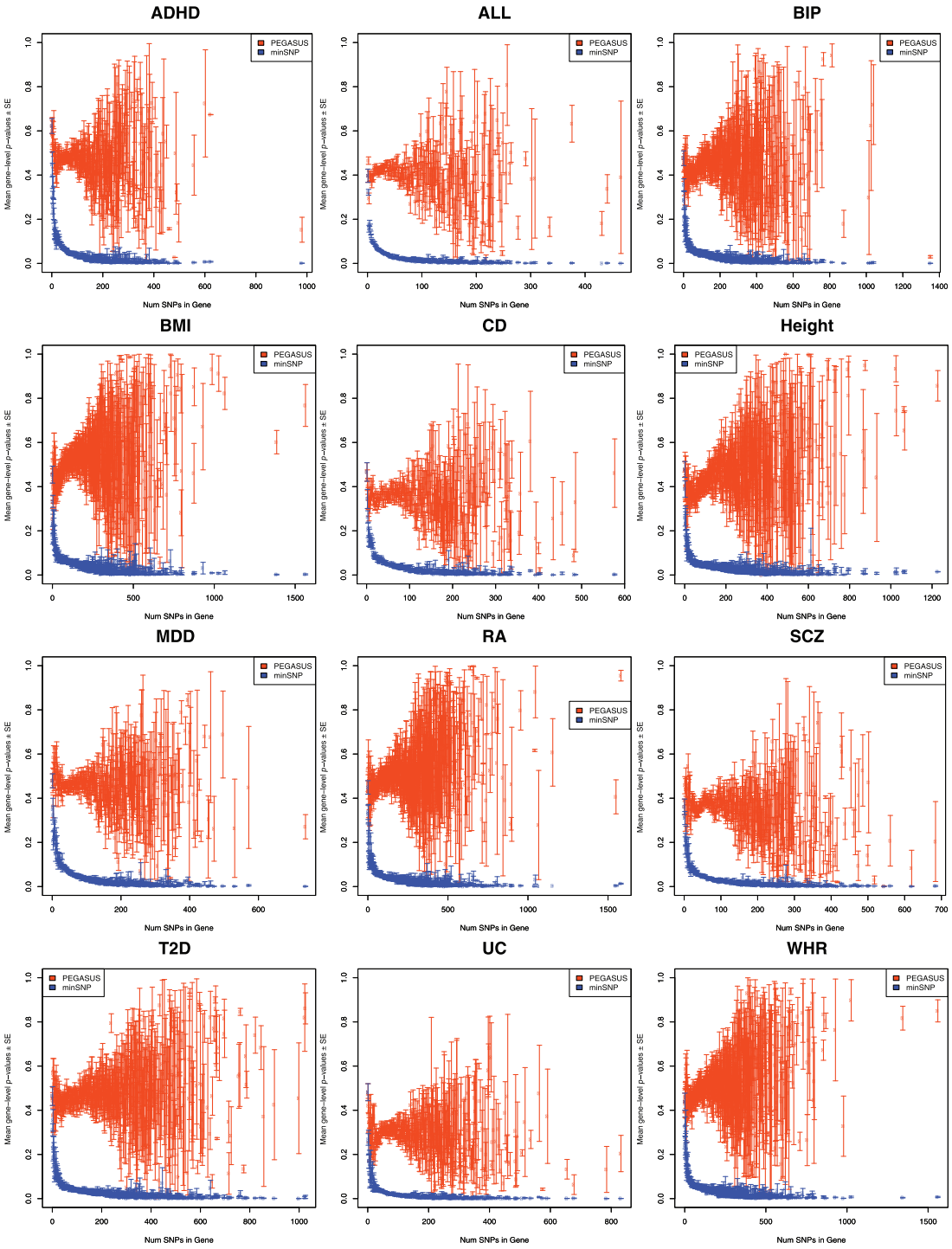
Receiver Operating Characteristic (ROC) Curves from GWA simulations Conducted with Varying Parameters. We performed a GWA study for a simulated phenotype with known underlying true causal genes (see Subjects and Methods), and determined true positive rate (TPR; genes truly associated with phenotype that were identified as such) and false positive rate (FPR; genes identified as causal by a gene score method that were not truly associated with the simulated phenotype) for minSNP, VEGAS, SKAT and PEGASUS for various gene score thresholds (see Subjects and Methods). A-C show ROC curves for simulations with 20%, 40%, and 60% spurious GWA hits, respectively, from real GWA data for Crohn's Disease (CD) added to GWA p-values for a simulated phenotype. D-F show ROC curves for simulations with 20%, 40%, and 60% spurious GWA hits, respectively, from real GWA data for Rheumatoid Arthritis (RA) added to GWA p-values for a simulated phenotype. H-J show ROC curves for simulations with 20%, 40%, and 60% spurious GWA hits, respectively, from real GWA data for Ulcerative Colitis (UC) added to GWA p-values for a simulated phenotype. We find that PEGASUS and VEGAS outperform minSNP and SKAT with higher TPRs at very low FPRs for every simulation.

Figure S15



Quantile-quantile plots comparing SKAT and PEGASUS p-values We calculated gene-level p-values for the Acute Lymphoblastic Leukemia (ALL) (Xu *et al.* (2013)) and the WTCCC Type 2 Diabetes (T2D) (WTCCC (2007)) datasets using the SKAT method with suggested weights ($\overline{w_j}$ = Beta(MAF$_j$; 1, 25) and no weights (w_j = 1 for all variants) (Wu *et al.* (2010, 2011)) and PEGASUS; we note that SKAT requires genotype data as input, which is why these two phenotypes were analyzed here. We show the resulting log_{10} PEGASUS p-values versus log_{10} SKAT p-values for the two datasets. For both ALL and T2D, we find that SKAT p-values calculated with the suggested weights and PEGASUS p-values are correlated, with r values of 0.4445 and 0.4933, respectively ($p < 2 \quad 10^{-16}$ for both datasets). When the unweighted SKAT test is performed, we find that SKAT p-values and PEGASUS p-values are correlated, with r values of 0.9609 and 0.9406, respectively ($p < 2 \quad 10^{-16}$ for both datasets).

Figure S16



Mean minSNP and PEGASUS gene scores versus gene length for the 12 datasets analyzed in this study (Table 2) Each point represents the mean gene score ± standard error of the mean (SE) for a given number of SNPs in a gene. We find that the mean minSNP gene score (blue) decreases with increasing number of SNPs in a gene, however, PEGASUS gene scores (orange) do not show this trend.

Table S1

Disease or Trait	Level of Significance for gene score (10^{-6})	# of significant genes (minSNP)	% significant genes (minSNP)	# of significant genes (PEGASUS)	% significant genes (PEGASUS)	# of significant genes (VEGAS)	% significant genes (VEGAS)
ADHD (NEALE *et al.* (2010))	2.81	1	0.01	0	0	0	0
ALL (XU *et al.* (2013))	2.78	29	0.16	10	0.06	10	0.06
BIP (SKLAR *et al.* (2011))	2.81	47	0.26	17	0.1	17	0.1
BMI (SPELIOTES *et al.* (2010))	2.83	96	0.54	56	0.32	56	0.32
CD (FRANKE *et al.* (2010))	2.88	431	2.48	256	1.47	249	1.43
Height (LANGO ALLEN *et al.* (2010))	2.83	834	4.71	404	2.28	397	2.24
MDD (RIPKE *et al.* (2013))	2.81	16	0.09	0	0	0	0
RA (STAHL *et al.* (2010))	2.80	252	1.41	180	1.01	183	1.03
SCZ (RIPKE *et al.* (2011))	2.81	250	1.41	55	0.31	55	0.31
T2D (MORRIS *et al.* (2012))	2.82	54	0.3	11	0.06	10	0.06
UC (ANDERSON *et al.* (2011))	2.80	980	5.5	190	1.07	191	1.07
WHR (HEID *et al.* (2010))	2.83	50	0.28	17	0.1	18	0.1

Number of significant genes and percentage of significant genes (out of the total number of genes considered) for minSNP, PEGASUS and VEGAS gene scores. Significance is based on Bonferroni correction for the total number of genes for which the study had SNP-level p-values in each of 12 GWA datasets (Table 2). We find that minSNP always finds more associated genes than the other two methods. PEGASUS and VEGAS typically find similar numbers of associated genes.

Table S2

Disease or Trait	Reference Data	# WTCCC Cases in Reference Data	# Non-WTCCC Cases in Reference Data	# of PEGASUS Hits in Reference Data replicated in WTCCC data	Total # PEGASUS hits in Reference Data	% PEGASUS hits replicated in WTCCC data (Column 5/Column 6)
BIP	SKLAR *et al.* (2011)	1,571	5,910	0	17	0.0
CD	FRANKE *et al.* (2010)	1,747	4,586	11	254	4.3
RA	STAHL *et al.* (2010)	1,525	4,014	103	180	57.2
T2D	MORRIS *et al.* (2012)	1,924	10,247	1	11	9.1

Replication of PEGASUS gene hits in WTCCC datasets. We conducted a replication experiment for PEGASUS gene hits for the 12 GWA datasets in this study in WTCCC data for Bipolar Disorder (BIP), Crohn's Disease (CD), Rheumatoid Arthritis (RA) and Type 2 Diabetes (T2D). A "hit" is a gene with a gene score $p_g < 2.8 \quad 10^{-6}$; this threshold is based on Bonferroni correction for the total number of genes for which the study had SNP-level p-values in each of 12 GWA datasets (Table 2). We find that we are able to replicate up to 57.2% of gene hits in the WTCCC dataset using the PEGASUS method. We note that these four meta-analyses included the WTCCC dataset and are comprised of much larger sample sizes than our replication dataset.

Table S3

Source of GWA p-values	**URL for Downloading p-values**
Psychiatric Genomics Consortium	https://www.med.unc.edu/pgc/downloads
the International IBD Genetics Consortium	http://www.ibdgenetics.org/downloads.html
GIANT consortium	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
Broad Institute	http://www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_etal_2010NG/
DIAGRAM Consortium	http://diagram-consortium.org/downloads.html
Acute Lymphoblastic Leukemia (ALL) data	Genotype data are available from dbGaP. (http://www.ncbi.nlm.nih.gov/gap), and steps to perform the GWA study are outlined in Xu et al (XU *et al.* (2013))

URLs for full SNP-level p-values from GWA datasets analyzed.

Algorithm S1

Algorithm 1 permSNP

1: Calculate the observed gene-level test statistic for gene with n SNPs

$$Q_{obs} = \frac{\sum_{i=1}^{n} q_i}{n} \tag{10}$$

2: For M case-control permutations, calculate the permuted gene-level test statistic and record whether the permuted statistic is greater than the observed statistic

3: **for** $j = 1$ to M **do**

4:

$$Q_j^* = \frac{\sum_{i=1}^{n} q_i^*}{n} \tag{11}$$

$$I_j = \begin{cases} 1 & \text{if } Q_j^* > Q_{obs} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

5: **end for**

6: Calculate gene-level p-value

$$p_g = \frac{\sum_{j=1}^{M} I_j}{M} \tag{13}$$

Text S1

Connection between SKAT and PEGASUS tests

The underlying model for the SKAT test is a multiple linear (or logistic, depending on the phenotype of interest) mixed-model regression, given by Eq. 2; the associated variance component score statistic is given by Eq. 3. Under the null hypothesis of no association, the SKAT model reduces to the following:

$$y_i = \alpha_0 + \boldsymbol{\alpha}' \boldsymbol{C_i} + \epsilon_i \tag{8}$$

In matrix notation, Eq. 8 can be written in matrix notation as $\mathbf{Y} = \boldsymbol{\alpha}'\mathbf{C} + \boldsymbol{\epsilon}$, where $\epsilon \sim N(0, \sigma^2)$ and $\mathbf{C}$ is a matrix of covariates, and the estimates of the regression coefficients can be expressed as $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + (\mathbf{C^T C})^{-1}\mathbf{C^T}\boldsymbol{\epsilon}$. We define $\mathbf{P} = \mathbf{C}(\mathbf{C^T C})^{-1}\mathbf{C^T}$, which is a projection matrix as is $\mathbf{I} - \mathbf{P}$. We substitute these values into the SKAT test statistic (Eq. 3), which gives the following:

$$\begin{aligned} Q &= [(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}]^T \boldsymbol{K}[(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}] \\ &= \boldsymbol{\epsilon^T}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon} \\ &= \boldsymbol{\epsilon^T}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{K}\boldsymbol{\epsilon} \end{aligned} \tag{9}$$

Let $d = d_1, \ldots, d_n$ be the eigenvalues of $\sigma^2[(\mathbf{I} - \mathbf{P})\mathbf{K}]$. Then, the test statistic takes the form of $\epsilon^{\mathbf{T}} diag(d_1, \ldots, d_n)\epsilon$; this is a quadratic form that follows a mixture of χ^2 distributions with weights given by d.

In PEGASUS, the null distribution is a mixture of χ^2 distributions with weights given by λ, the eigenvalues of the LD matrix Σ. If no covariates are considered and the variant weights are uniform for all variants ($w_j = 1$), the SKAT null distribution becomes a mixture of χ^2 distributions with mixture proportions given by the eigenvalues of the $\mathbf{K} = \mathbf{GWG'}$ matrix, which is a variance-covariance matrix similar to the PEGASUS LD matrix $\boldsymbol{\Sigma}$. Thus, under these circumstances, the two tests give similar results (Figure S15C-D).

Text S2

Additional significantly associated gene subnetworks for Attention-Deficit/Hyperactivity Disorder (ADHD), identified using PEGASUS gene scores as input to HotNet2

The subnetwork shown in Figure 5E contains *FURIN*, a gene containing one SNP significantly associated with ADHD (GWA p-value 1.316×10^{-5}; NEALE *et al.* (2010)). *FURIN* encodes the protease furin that processes latent precursor proteins into biologically active truncated BDNF molecules in the trans-Golgi network. Decreased levels of truncated BDNF have been shown to be associated with memory loss and learning impairment (CARLINO *et al.* (2013)). Another gene in the subnetwork, *PACS1*, has been found to be moderately associated with years of education in a previous GWA study (p-value 4.9×10^{-6}; RIETVELD *et al.* (2013)). PACS1, encoded by *PACS1*, is a trans-golgi-membrane traffic regulator that binds furin; it has been shown to be involved in the localization of the protease furin to the trans-Golgi network (WAN *et al.* (1998)). This gene has also been found to be mutated in cases of an unknown syndrome with intellectual disability (SCHUURS-HOEIJMAKERS *et al.* (2012)). LRP1, or low density lipoprotein receptor-related protein 1, is encoded by the *LRP1* gene and is a major receptor for apolipoprotein E (apoE), which transports lipids to the brain and also plays a role in neuronal repair (FUENTEALBA *et al.* (2010)). LRP1 and apoE are also involved in mediating the clearance of amyloid β 42 from the brain, and defective clearance of amyloid β leads to accumulation in neurons, which contributes to Alzheimer's Disease (FUENTEALBA *et al.* (2010)). Collectively, we find that genes in this subnetwork play important roles in cell trafficking processes that may be involved in the pathology of neuropsychiatric and cognitive disorders such as Alzheimer's Disease and intellectual disability.

A third subnetwork (Figure 5F) contains the genes *RORB* and *MPPED2*, both of which are moderately associated with years of education in previous GWA studies (GWA p-values 7.5×10^{-4} and 6.7×10^{-4}, respectively; RIETVELD *et al.* (2013)). *MPPED2* is also associated with information processing speed (LUCIANO *et al.* (2011): GWA p-value 1.6×10^{-5}). *MPPED2* is expressed in the fetal brain and is associated with WAGR syndrome whose symptoms include Wilms' tumor, aniridia, genitourinary anomalies and retardation, but the gene's function in the brain is still unclear (CHEN *et al.* (2010); DAVIS *et al.* (2008)). The gene *NR2F2*, which encodes the COUP-TFII transcription factor, is important in many aspects of neural development including neurogene-

sis, axogenesis, and differentiation (NAKA *et al.* (2008)). *RORB* encodes the transcription factor RORB which is known to play a regulatory role in neurogenesis and has been found to be associated with bipolar disorder (MCGRATH *et al.* (2009)). The gene *RARG* encodes a retinoic acid receptor (retinoic acid receptor gamma) which makes a heterodimeric pair with a retinoid X receptor. The RXRG gene encodes one such retinoid X receptor (MADEN (2007)). The retinoic acid receptor-retinoid X receptor heterodimeric pair are involved in mediating response to retinoic acid by binding to a DNA region containing a retinoic acid response element (RARE), which is bound to a corepressor protein and regulates transcription. Retinoic acid is very important in development of the nervous system, especially in patterning and neuronal differentiation (MADEN (2007)). The genes in this subnetwork (Figure 5F) are thus likely involved in brain development. The genes in Figure 5D-F were found using VEGAS gene scores as input to HotNet2 as well.

Text S3

Additional significantly associated gene subnetworks for Waist-Hip Ratio (WHR), identified using PEGASUS gene scores as input to HotNet2

The second subnetwork found by HotNet2 (LEISERSON *et al.* (2015)) using PEGASUS gene scores as input contains the gene *VEGFA* (vascular endothelial growth factor A), which is thought to mediate adipogenesis (HEID *et al.* (2010)). Serum levels of VEGFA are correlated with obesity (HEID *et al.* (2010)). VEGFA is also associated with angiogenesis (formation of new blood vessels) and for these reasons, VEGFA has been posited to underlie the association between childhood obesity and increased risk for atherosclerosis (SIERVO *et al.* (2012)). The growth factor encoded by *VEGFB* (vascular endothelial growth factor B), which is also contained in the subnetwork, is involved in endothelial targeting of fatty acids to peripheral tissues. Mice deficient in *VEGFB* show decreased uptake of lipids by heart, muscle and brown adipose tissue, which are mitochondria rich and use fatty acids as an energy source, and increased accumulation of lipids in white adipose tissue instead, ultimately resulting in increased body weight (HAGBERG *et al.* (2010)). *FLT1*, which encodes the vascular endothelial growth factor receptor 1, and NRP1 or neurophilin 1, encoded by the *NRP1* gene, are a receptor and co-receptor, respectively for VEGFB. Mice lacking the *FLT1* gene and mice deficient in *NRP1* both show down-regulation of fatty acid transport proteins in the heart, suggesting that VEGFB functions via FLT1 and NRP1 signaling (HAGBERG *et al.* (2010)). In this subnetwork, we see genes known to be associated with obesity and lipid trafficking; taken together, this result elucidates the mechanisms of action of VEGFA, which contains a known GWA study association for WHR (SNP p-value 1.38×10^{-10}; HEID *et al.* (2010)) and VEGFB, which is moderately associated with WHR (SNP p-value 7.2×10^{-3}; HEID *et al.* (2010)). These genes were found using VEGAS gene scores as input to HotNet2 as well.