

RESEARCH ARTICLE

Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics

David Lamparter^{1,2}✉, Daniel Marbach^{1,2}✉, Rico Rueedi^{1,2}, Zoltán Kutalik^{1,2,3}‡*, Sven Bergmann^{1,2}‡*

1 Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland, **2** Swiss Institute of Bioinformatics, Lausanne, Switzerland, **3** Institute of Social and Preventive Medicine (IUMSP), Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland

✉ These authors contributed equally to this work.

‡ ZK and SB also contributed equally to this work.

* Sven.Bergmann@unil.ch; Zoltan.Kutalik@unil.ch.



OPEN ACCESS

Citation: Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S (2016) Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput Biol* 12(1): e1004714. doi:10.1371/journal.pcbi.1004714

Editor: Jennifer Listgarten, Microsoft Research, UNITED STATES

Received: June 16, 2015

Accepted: December 17, 2015

Published: January 25, 2016

Copyright: © 2016 Lamparter et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Download location of meta analysis results are given in [S2 Table](#). Meta analysis pertaining to the colaus data set is given at (<http://www2.unil.ch/cbg/index.php?title=PascalTestData>). The software is available (<http://www2.unil.ch/cbg/index.php?title=Pascal>).

Funding: The CoLaus study was and is supported by research grants from GlaxoSmithKline (<https://www.gsk.com/>), the Faculty of Biology and Medicine of Lausanne, and the Swiss National Science Foundation (<http://www.snf.ch>) (grants 33CSCO-122661, 33CS30-139468 and 33CS30-148401). ZK received financial support from the Leenaards

Abstract

Integrating single nucleotide polymorphism (SNP) p-values from genome-wide association studies (GWAS) across genes and pathways is a strategy to improve statistical power and gain biological insight. Here, we present *Pascal* (Pathway scoring algorithm), a powerful tool for computing gene and pathway scores from SNP-phenotype association summary statistics. For gene score computation, we implemented analytic and efficient numerical solutions to calculate test statistics. We examined in particular the sum and the maximum of chi-squared statistics, which measure the strongest and the average association signals per gene, respectively. For pathway scoring, we use a modified Fisher method, which offers not only significant power improvement over more traditional enrichment strategies, but also eliminates the problem of arbitrary threshold selection inherent in any binary membership based pathway enrichment approach. We demonstrate the marked increase in power by analyzing summary statistics from dozens of large meta-studies for various traits. Our extensive testing indicates that our method not only excels in rigorous type I error control, but also results in more biologically meaningful discoveries.

Author Summary

Genome-wide association studies (GWAS) typically generate lists of trait- or disease-associated SNPs. Yet, such output sheds little light on the underlying molecular mechanisms and tools are needed to extract biological insight from the results at the SNP level. Pathway analysis tools integrate signals from multiple SNPs at various positions in the genome in order to map associated genomic regions to well-established pathways, i.e., sets of genes known to act in concert. The nature of GWAS association results requires specifically tailored methods for this task. Here, we present *Pascal* (Pathway scoring algorithm), a tool that allows gene and pathway-level analysis of GWAS association results without the need

Foundation (<http://www.leenaards.ch/>), the Swiss Institute of Bioinformatics (<https://www.isb-sib.ch/>) and the Swiss National Science Foundation (31003A-143914, 51RTP0_151019). SB received funding from the Swiss Institute of Bioinformatics, the Swiss National Science Foundation (grant FN 310030_152724 / 1) and SystemsX.ch through the SysGenetiX project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

to access the original genotypic data. *Pascal* was designed to be fast, accurate and to have high power to detect relevant pathways. We extensively tested our approach on a large collection of real GWAS association results and saw better discovery of confirmed pathways than with other popular methods. We believe that these results together with the ease-of-use of our publicly available software will allow *Pascal* to become a useful addition to the toolbox of the GWAS community.

Introduction

Genome-wide association studies (GWAS) have linked a large number of common genetic variants to various phenotypes. For most common phenotypes, high-powered meta-analyses have revealed tens to hundreds of single nucleotide polymorphisms (SNPs) with robust associations. However, deriving biological knowledge from these associations is often challenging [1,2]. Many genes function in multiple biological processes and it is typically not clear which of these processes is related to the phenotype in question.

Pathway analysis aims to provide insight into the biological processes involved by aggregating the association signal observed for a collection of SNPs into a pathway level signal. This is generally carried out in two steps: first, individual SNPs are mapped to genes and their association p-values are combined into gene scores; second, genes are grouped into pathways and their gene scores are combined into pathway scores. Existing tools vary in the methods used for each step and the strategies employed to correct for correlation due to linkage disequilibrium.

SNPs are usually mapped to genes based on physical distance [3], linkage disequilibrium (LD) [4], or a combination of both [5]. Genes are commonly assigned to pathways using well-established databases (such as Gene Ontology [6], KEGG [7], PANTHER [8], REACTOME [9], BIOCARTA [10]) or in-house annotation (based on co-expression [4], for example).

Various methods have been developed to aggregate SNP summary statistics into gene scores [3,11,12]. A common aggregation method is to use only the most significant SNP within a window encompassing the gene of interest, for example by assigning the maximum-of-chi-squares (MOCS) as the gene score statistic [3,13] (the contributing chi-squared values can be obtained from SNP p-values by using the inverse chi-squared quantile transformation). Another method is to combine results for all SNPs in the gene region, for example by using the sum-of-chi-squares (SOCS) statistic [14]. Both the MOCS and SOCS statistics are confounded by several properties of the gene. Specifically, in both cases it is important to correct for gene size and LD structure to obtain a well-calibrated p-value for the statistic. In the remainder of this paper, we also refer to the p-values of the MOCS and the SOCS statistics as *max* and *sum gene scores*, respectively. P-values can be estimated by phenotype label permutation, but this method is both computationally intensive and requires access to genotype data of the actual study, which are rarely shared [15]. Thus, one often has access only to association summary statistics and not the individual genotypes. In this case, one method is to regress out confounding factors [3]. This approach is employed in the popular MAGENTA tool, but provides only a partial solution as substantial residual confounding still remains [3].

An alternative approach, which we take here, is to exploit the fact that the null distributions of the MOCS and SOCS statistics depend solely on the pairwise correlation matrix of the contributing genotypes. In the absence of the original genotypes, this correlation matrix can still be estimated from ethnicity-matched, publicly available genotypic data, as has been proposed by us and others for conditional multi-SNP analysis of GWAS results [16,17]. This approach has

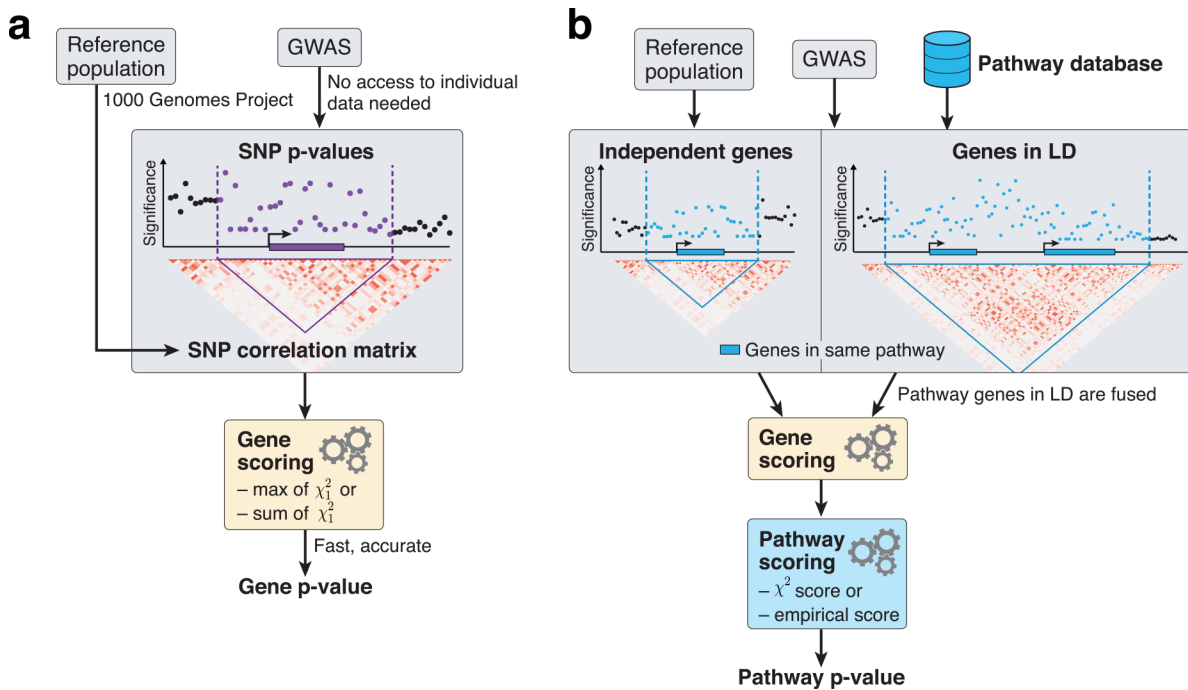


Fig 1. Overview of the methodology to compute gene and pathway scores. a) We compute gene scores by aggregating SNP p-values from a GWAS meta-analysis (without the need for individual genotypes), while correcting for linkage disequilibrium (LD) structure. To this end, we use numerical and analytic solutions to compute gene p-values efficiently and accurately given LD information from a reference population (e.g. one provided by the 1000 Genomes Project[22]). Two options are available: the max and sum of chi-squared statistics, which are based on the most significant SNP and the average association signal across the region, respectively. b) We use external databases to define gene sets for each reported pathway. We then compute pathway scores by combining the scores of genes that belong to the same pathways, i.e. gene sets. The fast gene scoring method allows us to dynamically recalculate gene scores by aggregating SNP p-values across pathway genes that are in LD and thus cannot be treated independently. This amounts to fusing the genes and computing a new score that takes the full LD structure of the corresponding locus into account. We evaluate pathway enrichment of high-scoring (possibly fused) genes using one of two parameter-free procedures (chi-squared or empirical score), avoiding any p-value thresholds inherent to standard binary enrichment tests.

doi:10.1371/journal.pcbi.1004714.g001

been implemented in the *Versatile Gene-based Association Study* (VEGAS) software and yields results close to those from phenotype label permutation[11]. However, while VEGAS is faster than estimation via phenotype label permutation, it still relies on a Monte Carlo method for estimating the p-values. This limits its efficiency for highly significant gene scores.

Once gene scores have been computed, pathway analysis tools use various strategies to aggregate them across sets of related genes. The most common approach used for analysing GWAS meta-analysis results, as exemplified by the popular GWAS pathway analysis tool MAGENTA, is based on binary enrichment tests, which rely on a threshold parameter to define which genes are significantly associated with the trait[3,18]. However, with this strategy potential contributions of weakly associated genes that just missed the threshold are lost and there is no clear guidance on how the threshold parameter should be set. Indeed, it seems common practice to keep the default parameter without knowing whether other choices would produce better results[5].

In this work, we focus on improving two major aspects of pathway enrichment analysis (Fig 1). First, we incorporate numerical and analytic solutions for the p-value estimation of the MOCS and SOCS statistics. This removes the need for phenotype permutations or Monte Carlo simulations, thereby making the score computation faster. Second, we developed a rigorous type I error control strategy and implemented a modified Fisher method to compute parameter-free pathway scores[19]. While some elements of our algorithm have been proposed

in other fields of statistical genetics[20,21], the novelty of our method lies in the unique combination of sophisticated analytical methods employed for pathway analysis, which results in improved computational speed, precision, type I error control and power.

In the following, we first evaluate the performance of our tool, demonstrating its speed gains and robust control of type I error. Then, using precision-recall analyses, comparing small to large GWAS results for lipid traits and Crohn’s disease, we demonstrate that our pathway scoring approach exhibits a gain in power compared to binary enrichment. Finally, we apply our method to dozens of large meta-analysis studies and evaluate power by counting the number of pathways passing the Bonferroni-corrected p-value threshold.

We provide this tool for gene and pathway scoring as a standalone, open-source software package called *Pascal*.

Results

Pascal computes genes scores rapidly and to very high precision

First, we compared the run time and precision of *Pascal* to those of VEGAS[11], one of the current state-of-the-art gene scoring tools. To this end, we applied both procedures to genome-wide p-values obtained from two large-scale GWAS meta-analyses: The first used about 2.5 million HapMap imputed SNPs[23,24] and the second was based on about 6.4 million SNPs imputed based on a common subset of 1000 Genomes Project (1KG) panel[22,25]. As benchmark we used the results from VEGAS for the former and VEGAS2 (a recent implementation of VEGAS that uses pre-computed LD matrices from 1KG[26]) for the latter. We observed a substantially smaller run time for our method in both cases (Fig 2A): for the HapMap imputed data, VEGAS took 29 hours to compute 18,132 gene scores, while *Pascal* was considerably faster, needing only about 30 minutes for either statistical test (sum or max) on a single core (Intel Xeon CPU, 2.8GHz). For the 1KG imputed data, *Pascal* finished the computation in under two hours for either statistic, whereas VEGAS2 took over ten days.

To compare the gene scores computed by the two methods, we increased the maximum number of Monte Carlo runs for VEGAS to 10^8 , at a high computational cost (about 9 days of runtime). We observed excellent concordance between the gene scores of *Pascal* and VEGAS,

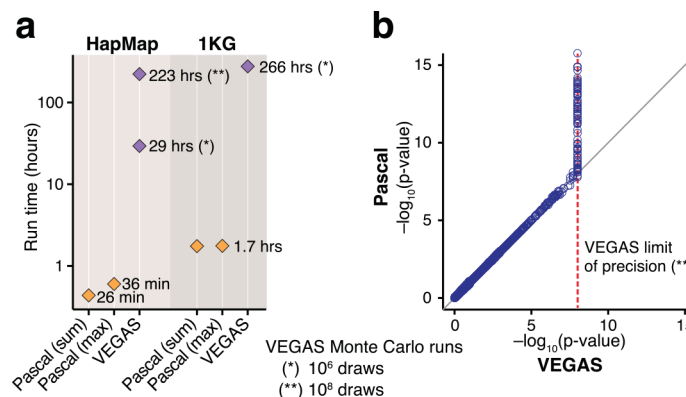


Fig 2. Comparing efficiency between VEGAS and *Pascal*. a) Run times of VEGAS and *Pascal* (both options). Gene scores were computed on two GWAS (one HapMap imputed[23], one 1KG imputed[22,25]) for 18,132 genes on a single core. *Pascal* was compared to VEGAS for the HapMap imputed study and VEGAS2 for the 1KG-imputed study. For this plot, VEGAS and VEGAS2 were used with the default maximum number of Monte Carlo samples of 10^6 for both studies and additionally with 10^8 Monte Carlo samples for the HapMap imputed study. b) Scatter plot of $-\log_{10}$ -transformed gene p-values for the sum gene scores obtained by VEGAS and *Pascal*, respectively. P-values above 10^{-6} are in excellent concordance. Below this value VEGAS could not give precise estimates, since it was run with the maximal number of Monte Carlo samples set to 10^6 .

doi:10.1371/journal.pcbi.1004714.g002

except for scores below 10^{-6} : since we restricted VEGAS to 10^8 Monte Carlo runs, it could not estimate p-values smaller than 10^{-6} with good precision (Fig 2B). In contrast, *Pascal* can compute gene scores with high precision for p-values down to 10^{-15} . In summary, the analytic solutions incorporated in the *Pascal* algorithm offer a dramatic increase in efficiency and precision. Direct comparison of the sum and max gene scores of *Pascal* revealed good concordance between the two scoring methods. In cases where the results of two methods disagree, max scores tend to be more significant (S3 Fig).

The results reported here are all based on GWAS of European cohorts, thus we used the European panel from 1KG as reference panel. To evaluate whether this panel approximates LD matrices derived from other European cohorts sufficiently well, we compared results when using genotypes taken from the *CoLaus* cohort as reference panel[27]. We saw good concordance between the different reference panels for both the sum and the max gene scores for the largest HDL blood lipid GWAS to-date[23] (S2 Fig).

Pascal controls for inflation due to neighbouring genes

In general, methods that compute pathway scores from gene scores assume independence of these scores under the null hypothesis. However, neighbouring genes often have correlated scores due to LD, and are sometimes part of the same pathway. This results in a non-uniform pathway score p-value distribution under the null hypothesis. MAGENTA deals with this problem by pruning gene scores based on LD and using only the highest gene score in the region. However, this introduces a bias toward high gene scores into the calculation of pathway scores[3].

Our fast gene score calculation allows us to address this issue with a gene-fusion strategy. In brief, for each pathway harbouring correlated genes, gene scores are recomputed jointly for each correlated gene set (i.e. fusion-gene) using the same method as for individual genes (Fig 1B, Methods), thus taking the full LD structure of the corresponding region into account.

To see if our approach provides well-calibrated p-values, we simulated random phenotypes and calculated association p-values for all 1KG SNPs. We then employed our pathway analysis pipeline and checked if pathway p-values were uniformly distributed, as expected for random phenotypes. We found that without the gene-fusion strategy, pathway p-values are indeed inflated and, as expected, this inflation is stronger for pathways with many proximal genes (Fig 3A). In contrast, applying the gene-fusion strategy corrects the distribution of pathway score p-values to be uniform irrespective of the number of proximal genes (Fig 3B). Importantly, we did not see inflation for very small p-values with the gene-fusion strategy, which is essential for type I error control.

Going one step further, we also simulated *in-silico* phenotypes influenced by randomly selected causal SNPs. We explored two scenarios: one where 50 SNPs were randomly selected from the entire genome and another where random sampling was applied to gene regions only. The experiment was repeated 50 times and independent genetic data was used to generate the estimated pairwise correlation. Although in this case gene scores naturally deviate from the null distribution, we found that overall pathway p-values remain well calibrated (S14 and S15 Figs). Note that we explored only a limited set of simulation scenarios and cannot exclude that some settings might produce less well-calibrated results (see legend of S15 Fig).

Pascal has higher sensitivity and specificity than hypergeometric pathway enrichment tests

A commonly used statistic to derive pathway scores from a ranked list of genes (or SNPs) is to first apply a fixed threshold in order to define a subset of elements that is considered to be significantly associated with the given trait. The pathway statistic is then computed using a

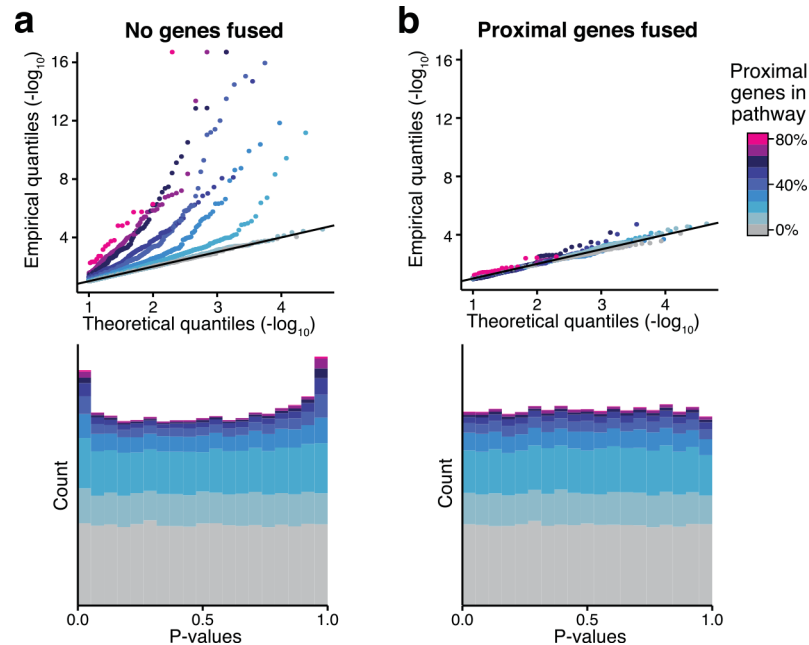


Fig 3. Pathway scores for random phenotypes. As input data we used 100 simulated instances of a random Gaussian phenotype and genotype data for 379 individuals from the EUR-1KG panel. Using the *Pascal* pipeline with sum gene scores and chi-squared pathway integration strategy we computed p-values for 1,077 pathways from our pathway library (results for max gene scores are similar, see [S4 Fig](#)). Panel (a) shows the p-value distributions without merging of neighbouring genes and (b) with merging of neighbouring genes (gene-fusion strategy). P-value distributions are represented by QQ-plots (upper panels) and histograms (lower panels). Results are colour-coded according to the fraction of genes in a given pathway that have a neighbouring gene in the same pathway, i.e. that are located nearby on the genome (distance <300kb). (a) P-values of pathways that contain genes in LD are strongly inflated without correction. (b) The gene fusion approach provides well-calibrated p-values independently of the number of pathway genes in LD.

doi:10.1371/journal.pcbi.1004714.g003

hypergeometric test evaluating whether the pathway contains more significant elements than expected. This approach is implemented, for example, in the tool MAGENTA[3]. Another common strategy is to use the rank-sum (Wilcoxon) test[3,28,29].

As described above, *Pascal* computes aggregate statistics without the need for defining a set of significant genes. We thus sought to compare this strategy with methods based on the hypergeometric or rank-sum tests. To this end, we tested performance on association results for four blood lipid traits obtained from of the *CoLaus* cohort[27]. We used a large meta-analysis of 188,577 individuals to define a reference set of associated pathways for each of the four lipid traits[23]. We then applied both pathway analysis methods to three non-overlapping, small subsets (1500 individuals) of the *CoLaus* study and compared how well the resulting pathways matched the reference set from the large study. We used the area under the precision-recall curve (AUC-PR) to quantify the performance of each method. Note that our choice was driven by the fact that precision-recall curves are preferred over receiver-operator-characteristic (ROC) curves when only a small fraction of tested pathways are in the reference set[30]. Our results show that *Pascal* outperforms both the hypergeometric and rank-sum based approaches (Fig 4A). Importantly, the better performance of *Pascal* is observed across a range of thresholds defining significant genes, including the optimal choice which is variable and *a priori* unknown across the different lipid phenotypes.

We applied the same evaluation strategy for GWAS data on Crohn's disease. We used the currently largest GWAS for Crohn's disease[31] to define a reference standard of associated

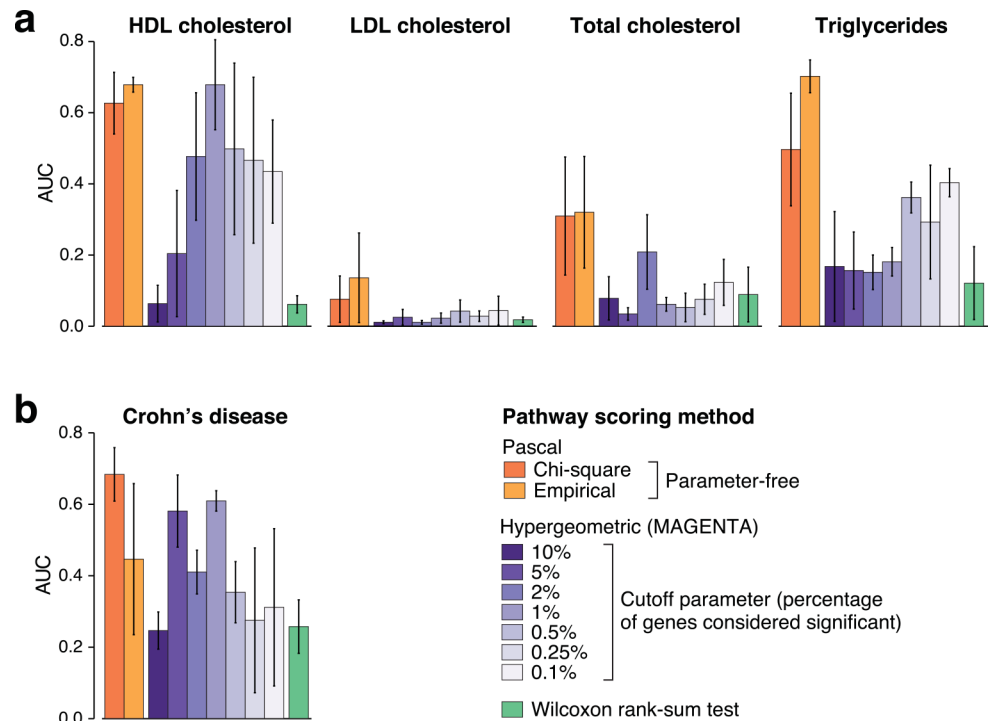


Fig 4. Performance of pathway enrichment methods for blood lipid traits and Crohn's disease. Displayed is the mean area under the precision-recall curve (AUC) for pathways identified using *Pascal*, a standard hypergeometric test at various gene score threshold levels, and a rank-sum test (vertical bars show the standard error). We show results for the max gene scores (sum gene score results are similar, see S5 Fig). a) Results for four blood lipid traits. The gold standard pathway list was defined as all pathways that show a significance level below 5×10^{-6} for any of the tested threshold parameters for hypergeometric tests in the largest study of lipid traits to date [23]. The significance level of 5×10^{-6} corresponds to the Bonferroni corrected, genome-wide significance threshold at the 0.5% level for a single method. For each phenotype, error bars denote the standard error computed from three independent subsamples of the *CoLaus* study (including 1500 individuals each). We see good overall performance of *Pascal* pathway scores, whereas results for discrete gene sets vary widely with the particular choice for the threshold parameter of hypergeometric test. b) Results for Crohn's disease using the same approach as in (a). A reference standard pathway list was defined as in (a) using the largest study of Crohn's disease traits to date [31]. We observe that the chi-squared strategy performs at least as well as all other strategies in this setting, whereas performance of the hypergeometric testing strategy varies.

doi:10.1371/journal.pcbi.1004714.g004

pathways. We then applied both pathway analysis methods to results from two individual cohorts participating in the meta-analysis that contained at least 1000 cases [31–33]. We observed that the chi-squared-method performed at least as well as all other strategies in this setting (Fig 4B). Overall, we saw similar results for both max and sum gene scores (S5 Fig).

Pascal has higher power than hypergeometric test based pathway enrichment in a wide range of traits

Having established that *Pascal* accurately controls type I error rate for simulated phenotypes and better recovers truly associated pathways for blood lipid traits as well as Crohn's disease, we next sought to evaluate its power when applied to large meta-analytic studies on a broad range of traits, where no ground truth can be defined.

To this end, we compared *Pascal* with the methods based on the hypergeometric test (using 9 different threshold values) and the rank-sum test proposed by Segrè *et al.* [3] for 118 GWAS

(S1 Table). All GWAS were derived from European populations justifying the use of the European 1KG genotypes as reference population. For a given GWAS, we asked how many tested pathways reached genome-wide significance at the Bonferroni-corrected p-value threshold of 0.05. Our results indicate that globally our approach has higher power than either the methods using the hypergeometric test (across all tested thresholds), or the rank-sum test (Figs 5A and S6). For individual traits (Fig 5B), specific choices of the threshold parameter of the hypergeometric test sometimes reveal more pathways, but again the value of the optimal threshold varies across traits and cannot be known *a priori*.

When splitting the GWAS into high powered (more than 50,000 individuals) and low powered studies (less than 50,000 individuals), we saw that in both cases we gain power by using *Pascal* although the effect was more pronounced for low powered GWAS (S7 Fig).

Hypergeometric enrichment testing is hampered by the fact that the optimal threshold is not known in advance. A strategy to overcome this could be to merge hypergeometric pathway scores coming from different sets of thresholds, further corrected for the effective size of the threshold sets. While such an aggregated hypergeometric testing improved performance, it was still outperformed by *Pascal* (S10 and S11 Figs).

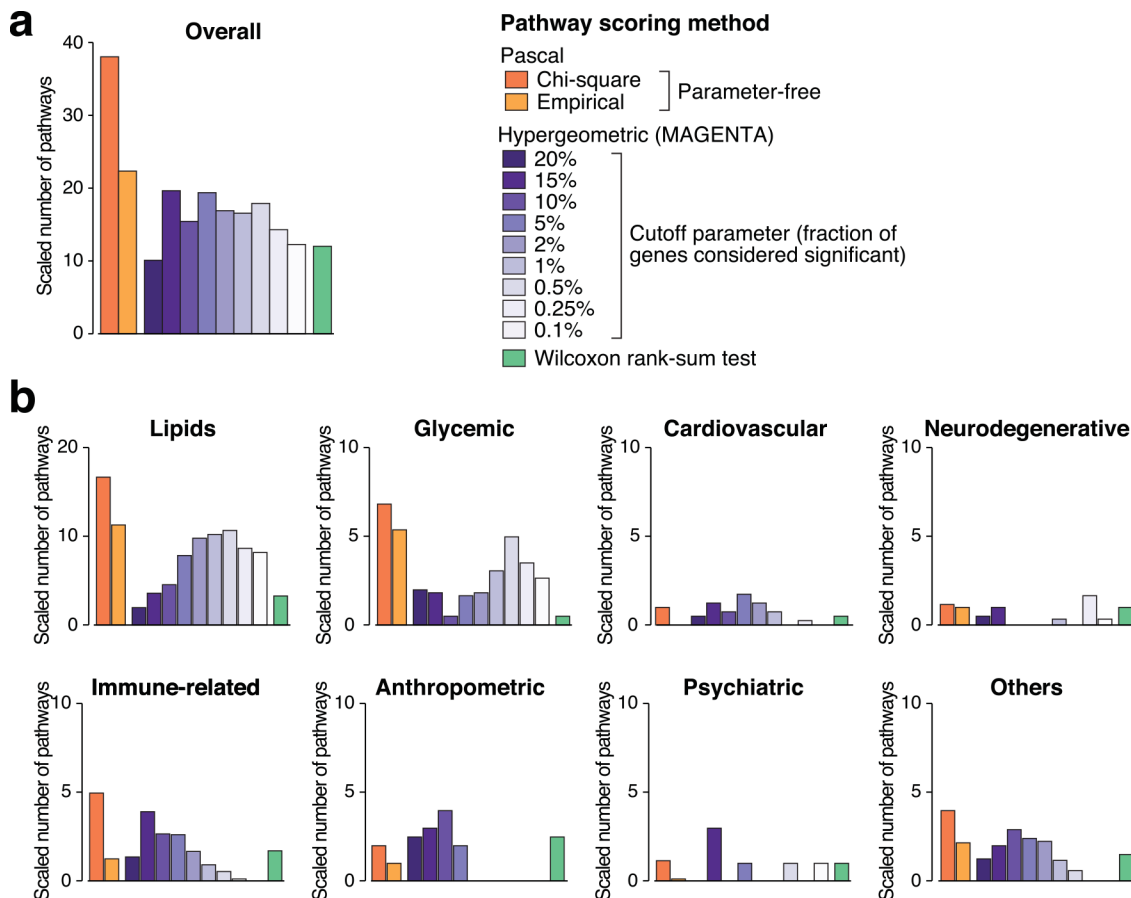


Fig 5. Power of pathway scoring methods across diverse traits and diseases. Bar heights represent the number of pathways found to be significant after Bonferroni-correction. Within a given trait group, results are aggregated for all tested GWAS studies. 65 GWAS had at least one significant pathway in one of the tested methods. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. This was done in order to avoid that a few studies with many emerging pathways dominate. We show results for the MOCS gene scores (SOCS gene score results are similar, see S6 Fig). (a) Results are aggregated over all trait groups. (b) Results for different trait groups.

doi:10.1371/journal.pcbi.1004714.g005

One of the proposed pathway scoring methods transforms the ranked gene p-values such that they follow a chi-squared distribution. The chi-squared distribution is a special case of the Gamma distribution with shape parameter 0.5. Thus we also examined whether using other shape parameters of the Gamma distribution could improve performance (see [Methods](#), [S12](#) and [S13](#) Figs). This analysis suggested that the chi-squared pathway scoring method represents a good compromise for a wide range of genetic architectures.

We found numerous examples of biologically plausible pathways discovered by *Pascal* that were not found by a standard binary enrichment analysis ([Fig 6](#), [S2 Table](#)). For insulin resistance[34] we found the REACTOME pathway *insulin signal attenuation* to be genome-wide significant. Notably, none of the genes in this pathway was found to contain a genome-wide significant SNP in the original publication. Another example is bone mineral density in women (LS-BMD)[35]. We found the *Hedgehog* and *Wnt* pathways to be significant, both of which are known to be involved in osteoblast biology[36]. Again, standard binary enrichment did not reach genome-wide significance. For smoking behaviour (measured in cigarettes per day)[37], we found pathways related to *nicotinic acetylcholine receptors*. For macular degeneration, we found *lipoprotein* and *complement system* involvement, which both have support in the literature[38,39]. These examples illustrate that the improvements made by *Pascal* not only lead to better performance on benchmarks, but may also have a dramatic impact on the interpretation of GWAS results in practice.

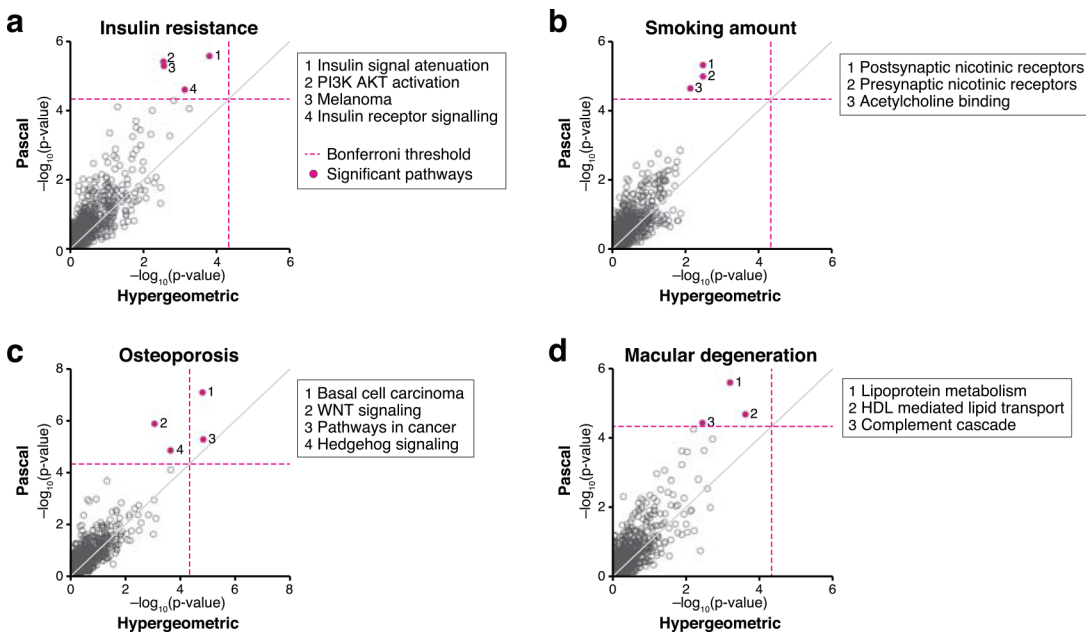


Fig 6. Examples of pathway enrichments comparing *Pascal* (chi-squared method) to the hypergeometric method. Displayed are results for four phenotypes showing improvement when using *Pascal* instead of the hypergeometric (binary) enrichment strategy at the 5% threshold level. Underlying gene scores were calculated using the sum method. Dashed lines refer to the Bonferroni significance level when correcting for the number of pathways (1077). Besides from few cancer-related pathways, all pathways highlighted by this analysis have been implied by prior research (see main text). (a) For the trait *insulin resistance*, *Pascal* scored the pathway *insulin signal attenuation* first, followed by two other trait-relevant pathways (*PI3K AKT activation* and *insulin receptor signaling*), while the hypergeometric test did not find any significant pathways. (b) For *smoking amount* (number of cigarettes per day), *Pascal* revealed three significant pathways related to *nicotinic acetylcholine receptors*. (c) For *osteoporosis*, two cancer-related pathways scored significant using both *Pascal* and the hypergeometric test, but only *Pascal* revealed the WNT and Hedgehog signaling pathways, which are known to be involved in osteoblast biology. (d) For *macular degeneration*, *Pascal* found three significant, trait-relevant pathways related to lipoproteins and the complement system.

doi:10.1371/journal.pcbi.1004714.g006

Discussion

In this work, we presented a new tool called *Pascal* (*Pathway scoring algorithm*) that specifically addresses both gene scoring and pathway enrichment, making significant advancement with respect to the state-of-the-art:

First, our gene score calculation combines analytical and numerical solutions to properly correct for multiple testing on correlated data[21]. While some of these approaches have already been applied within the rare variant field[20] (typically in a gene-wise fashion) we provide a streamlined implementation that can run genome-wide analyses without the need for any Monte Carlo simulations (making it about 100 times faster and more precise than the widely used software VEGAS).

Second, our pathway scoring integrates individual gene scores without the need for a tuneable threshold parameter to dichotomize gene scores for binary membership enrichment analysis (as done for example by MAGENTA). The choice of such a parameter is not straightforward and our method usually performs better, regardless of the chosen parameter.

Third, we show that the null distribution of enrichment p-values for pathways that contain genes in linkage disequilibrium is non-uniform due to an “over-counting” of gene association signals. This is a potential source of type I error underestimation and our method corrects for this phenomenon using a *gene fusion* approach, which considers genes in LD as single entities.

We have extensively evaluated the performance of *Pascal* for several real data sets. These comparisons demonstrated the rigorous control of type I error and superior predictive power in a wide range of trait and power settings in terms of enhanced precision-recall curves.

As an additional global measure of power, we considered the number of significantly enriched pathways for a large number of GWAS meta-analysis summary statistics. On average, our approach resulted in higher numbers of significant pathway scores than any binary enrichment strategy. Given its precise type I error control, this provides additional evidence of increased power for a wide range of traits. Indeed, the elevated rate of putatively involved pathways produced by our method not only reflects its higher sensitivity, but also already generates new hypotheses for further studies.

Taken together, our results demonstrate the superior performance of our approach compared to standard binary enrichment and rank-sum tests. Although methods with tuneable parameters might yield improved results in a particular setting, it is difficult to predict the optimal parameter choice. Indeed, the optimal parameter depends on sample size, as well as complexity and heritability of the phenotype. Another issue with binary enrichment tests is that the hypergeometric distribution is discrete, which leads to conservative p-values, especially if the expected number of successful draws is low. Our pathway scoring approach avoids this problem. Also, our approach lends itself to naturally extending pathway scoring in case genes have probabilistic membership in predefined pathways.

Users of our method will still have to make two choices: how to convert SNP p-values to gene scores (max or sum gene scores), and how to transform gene scores into pathway scores (empirical or chi-squared). We do not see evidence that one gene scoring method systematically outperforms the other in the context of our chi-squared pathway scoring method, while there seems to be a better performance for sum gene score when using the empirical approach (S8 Fig). To investigate this phenomenon we winsorized p-values (i.e. extreme p-values below 10^{-12} were set to 10^{-12}) and saw that the max gene score combined with empirical sampling suffered far less performance loss (S9 Fig). We therefore conclude that the power loss is due to outlier gene scores. The max gene-scores can lead to very high gene scores for high-powered studies. In the extreme case one gene might reach scores so high that it precludes detection of pathways not containing that gene when *the empirical sampling* strategy is used.

Future work could attempt to enhance several other aspects of our pathway enrichment analysis. For example, here we mapped SNPs to genes only based on physical distance, while potential improvements could be attained by incorporating additional information, such as eQTL data [40] and functional annotations, to assign weights to different association signals within a locus. While our approach is amenable to such a weighting scheme, this would potentially require the introduction of tuneable parameters, which we avoided so far. Furthermore, one may attempt to redefine gene sets based on external unbiased large-scale molecular data, such as expression data, while so far we only used the established (but likely biased) pathway collections [4]. To this end, we already integrated *Pascal* into a pipeline to analyze the connectivity between trait-associated genes across over 400 tissue-specific regulatory, co-expression and protein-protein interaction networks, further demonstrating its value for network-based analysis of GWAS results (Marbach *et al.*, submitted).

As an additional caveat, we should mention that *Pascal* uses the European 1KG sample as reference population per default. This choice may not be appropriate if the studied sample is not of European origin. In this case the user is encouraged to supply *Pascal* with the appropriate reference panel. Also, SNPs with low MAF are by default excluded from the analysis, because the low number of individuals in the European 1KG sample limits the accuracy of the LD estimate for low frequency variants. If the user wishes to include lower frequency variants, the use of a reference sample containing more individuals is recommended.

To conclude, *Pascal* implements fast and rigorous analytical methods into a single analysis pipeline tailored for gene scoring and pathway enrichment analysis that can be run on a desktop computer. We thus hope that *Pascal* will be useful to the GWAS community in a range of applications and play a pivotal role in leveraging the rich information encoded in GWAS results both for single traits and—given its efficiency and power—in particular also for high-dimensional molecular traits.

Our tool is available as a single standalone executable java package containing all required additional data at: <http://www2.unil.ch/cbg/index.php?title=Pascal> (short URL: <http://goo.gl/t4U5z6>).

Materials and Methods

Gene scores

The *Pascal* gene scoring method consists of the following steps (Fig 1A). First, we assign SNPs to genes if they are located within a given window around the gene body. For the experiments reported in this paper, we used windows extending 50kb up and downstream from the gene. A reference population is required to estimate the correlation structure between Z-scores of SNP association values. Here, we used the European population of the 1000 Genomes Project (1KG) [22], which allows us to apply our approach flexibly to summary statistics from diverse panels (HapMap, 1KG imputed, metaboChip or ImmunoChip).

Under the null hypothesis, it can be shown that the Z-scores of n SNPs in our gene region as multivariate normal:

$$\mathbf{z} \sim \mathcal{N}_n(0, \Sigma)$$

where Σ is the pair-wise SNP-by-SNP correlation matrix (see Section ‘Derivation of the sum score’ for details).

We define our base statistics, the SOCS (T_{sum}) and MOCS (T_{max}), as:

$$T_{sum} = \sum_{i=1}^n z_i^2$$

and

$$T_{max} = \max(z_i^2),$$

respectively. It can be shown that T_{sum} is distributed according to the weighted sum of χ_1^2 -distributed random variables:

$$T_{sum} \sim \sum_i \lambda_i \chi_1^2$$

where λ_i is the i -th eigenvalue of Σ . Its distribution function can be evaluated numerically (see Section [Algorithmic details for gene-score calculations](#) for details). To estimate the null distribution of T_{max} we make use of the fact that

$$P[T_{max} \geq t] = P[\max(|z_i|) \geq t] = 1 - P(|z_i| < t, i = 1, 2, \dots, n).$$

This amounts to a rectangular integration over a multivariate normal, for which an efficient algorithm is available[41]. The current implementation of this integration is suitable to estimate p-values larger than 10^{-15} . To approximate gene-wise p-values below this limit we multiply the minimum p-value of SNPs in the region with the effective number of tests within the gene (see Section [Algorithmic details for gene-score calculation](#)).

Gene fusion

Pathway analysis methods typically assume that the gene scores used to define pathway enrichment are independent. However, functionally related genes often cluster on the genome and harbor SNPs in LD, leading to correlated gene scores that violate this assumption. To circumvent this problem, we check for a given pathway if any of its genes that cluster physically close on the chromosome are in LD. If so, for the calculation of the pathway score, we consider a single entity (a so-called *fusion-gene*) consisting of all the SNPs of the gene cluster. We then replace the genes in the cluster by this *fusion-gene* and calculate its gene score, but only for the calculation of the score for this particular pathway. The pathway score is then computed from the p-values of independent pathway genes and fusion genes that integrate the associational signals from dependent pathway genes (Fig 1B). In this way, the LD structure of neighbouring pathway genes is taken into account. Our gene scoring method facilitates this approach because it is sufficiently fast and scalable for recomputing the scores of all fusion genes.

Pathway scores

For pathway analysis, we propose a parameter free enrichment strategy that does not require the specification of a gene score threshold, and thus allows weakly associated genes to contribute to pathway enrichment. The general approach consists of three steps: (1) gene scores are transformed so that they follow a target distribution, (2) a test statistic is computed by summing the transformed scores of pathway member genes and fusion-genes, and (3) analytic or empirical methods are used to evaluate whether the observed test statistic is higher than expected, i.e., the pathway is enriched for trait-associated genes. We considered two variants of this approach for pathway scoring (see overview in S1 Fig). The first variant is termed as the **chi-squared method**:

1. Gene score p-values are ranked such that the lowest p-value gets the highest rank. The rank value is then divided by the number of genes plus one to obtain a uniform distribution.
2. Uniform distribution values are transformed by the χ_1^2 -quantile function to obtain a χ_1^2 -distribution of gene scores.

3. χ_1^2 -gene scores of a given pathway of size m are summed and tested against a χ_m^2 -distribution.

The second variant is the *empirical sampling* method:

1. Gene score p-values are directly transformed with the χ_1^2 -quantile function to obtain new gene scores: $F_{\chi_1^2}^{-1}(1 - p)$.
2. A raw pathway score for a pathway of size m is computed by summing the transformed gene scores for all pathway genes.
3. A Monte Carlo estimate of the p-value is obtained by sampling random gene sets of size m and calculating the fraction of sets reaching a higher score than gene set of the given pathway.

We also tested a generalization of the chi-squared method where the inverse χ_1^2 -quantile transformation of the p-value ranks was replaced by the inverse Gamma-quantile transformation with varying shape parameter. For shape parameter of 0.5, the results coincide with results from the chi-squared method.

For our benchmarking procedures we created a pathway library by combining the results from KEGG[7,42], REACTOME[9] and BIOCARTA[10] that we downloaded from MsigDB [43].

Derivation of the sum-score

Let \mathbf{z} be the vector of Z-statistics coming from regressing the phenotype on each of the n SNPs within a gene-region. By construction, each Z-statistic has zero mean under the null. When both the outcome trait and the genotypes are standardized, the linear regression Z-statistics are essentially the scalar products of the genotype and the phenotype vectors. In other words, each Z-statistic in the region represents a weighted average of the same set of independent, identically distributed random variables. It can be shown that the correlation between two such mixtures, i.e. two Z-statistics, equals to the correlation between the weights, i.e. the correlation between the corresponding SNPs. Thus, the covariance matrix of \mathbf{z} is simply the pairwise SNP-by-SNP correlation matrix, denoted by Σ . Furthermore, the central limit theorem ensures that in case of sufficiently large sample size the Z-statistics are normally distributed. These facts put together yield that—under the null-hypothesis that no signal is present— \mathbf{z} follows a multivariate normal distribution, $\mathbf{z} \sim \mathcal{N}_n(0, \Sigma)$. For a detailed derivation see supplementary material in Xu et al[44] for example. Note that the between SNP correlation matrix Σ can be estimated from external data[17,45].

The eigenvalue decomposition of Σ is

$$\Sigma = \Gamma \Lambda \Gamma^T,$$

where Γ and Λ are the matrices of eigenvectors and eigenvalues, respectively. We see that multiplying \mathbf{z} with the inverse of the square-root of Σ leads to a vector of independent random variables. Let \mathbf{y} be defined as

$$\mathbf{y} = \Lambda^{-1/2} \Gamma^T \mathbf{z},$$

then

$$\mathbf{y} \sim \mathcal{N}_n(0, \mathbf{I}_n).$$

It follows that

$$\mathbf{z}^T \mathbf{z} = \mathbf{z}^T \mathbf{\Gamma} \mathbf{\Gamma}^T \mathbf{z} = \mathbf{y}^T \mathbf{\Lambda} \mathbf{y} \sim \sum_i \lambda_i \chi_1^2,$$

where λ_i is the i -th eigenvalue of Σ and χ_1^2 represents the chi-squared distribution.

Parameter settings

If not stated otherwise, our tool was always used with the following settings. We extended gene regions by 50kb upstream and downstream for gene scoring. Only SNPs that reached a MAF of 0.05 in European 1KG sample were used. For pathway score calculation, we removed the HLA-region. The gene-fusion parameter was set to 1Mb, so that when calculating a particular pathway score, all pathway-member genes less than 1Mb apart were fused for the calculation. We also removed genes containing more than 3000 SNPs except during speed benchmarking (Fig 2) where all SNPs were used.

Simulation settings for type I error control of the pathway scores

We used genotypes for 379 individuals from the EUR-1KG cohort [22]. Corresponding phenotype values were simulated as independent, standard normally distributed variables. Univariate Z-scores for each of the 2,692,429 tested SNPs were calculated using linear regression. Simulations were repeated 100 times. Since we investigated the impact of gene-fusion, the LD matrix was estimated from the same data set to avoid any influence that might come from out-of-sample LD estimation.

Algorithmic details for gene-score calculations

Max-score. The algorithm first tries to use Monte Carlo simulation to derive p-values. Should the p-value be too small to be estimated within a few Monte Carlo draws, the procedure makes use of an algorithm for rectangular multivariate normal integration [41]. The implementation of the integration algorithm that is used is suitable to estimate p-values larger than 10^{-15} . In addition, this implementation is limited to correlation matrices of size below 1000 due to numerical stability concerns. Therefore, SNPs that are in very high LD ($r^2 > 0.98$) are pruned to lower the size of the correlation matrix. If more than 1000 SNPs fall into the gene or the gene-wise p-value is below 10^{-15} , we approximate the gene score by multiplying the minimal SNP-wise p-value in the gene region by the effective number of tests. The effective number of tests is calculated as the minimum number of principal components needed to explain 99.5% of total variance [46].

Sum-score. The algorithm relies on the Davies algorithm to calculate distribution function values of weighted sums of independent χ_1^2 -distributed random variables [47]. In case of convergence problems the Farebrother algorithm is used as a backup [48,49].

Web resources. A stand-alone executable for *Pascal* can be found at <http://www2.unil.ch/cbg/index.php?title=Pascal>. The *Pascal* source code can be found at <https://github.com/dlampart/Pascal>.

Supporting Information

S1 Fig. Overview of pathway scoring strategies in *Pascal*. Pathway scores are computed from gene scores. The upper panel shows a typical gene score distribution, where the pathway gene scores are indicated in black. In order to compute pathway scores, the original gene score p-values need to be transformed. To this end we use one of two strategies: in our *empirical strategy* (lower left panel), gene score p-values are directly transformed with the inverse χ_1^2 -quantile

function $F_{\chi^2_i}^{-1}(1 - \rho)$ to obtain scores, which are then summed across all pathway genes. A Monte Carlo estimate of the p-value is then obtained by sampling random gene sets of the same size and calculating the fraction of sets reaching a higher score than that of the given pathway. In the *chi-squared method* (bottom right panel), the gene score p-values are first ranked such that the lowest p-value ranks highest. The rank values are then divided by the number of genes plus one to define new p-values (p_{rank}) that are distributed uniformly by definition. From there, we proceed as for the empirical strategy just replacing p by p_{rank} . Also, since the scores are guaranteed to be chi-squared distributed, the computation of their corresponding p-value can be done analytically without any loss in precision.

(PDF)

S2 Fig. Comparison of results for different reference panels. Comparing p-values computed using LD matrices from the European 1000 Genome reference panel and the *CoLaus* cohort. GWAS summary statistics were taken from a large-scale blood-HDL level meta-analysis. Results are compared for (a) max gene scores; (b) max gene scores excluding gene scores that were computed with the effective number of tests approximation; and (c) sum gene scores. There is good concordance in all cases.

(PDF)

S3 Fig. Comparison of max and sum gene scores. We compared max and sum gene scores directly for a large-scale blood HDL level meta-analysis. Only gene scores up to 10^{-15} are displayed, which truncated 6 genes with very large max scores. R^2 between the $-\log_{10}$ -transformed variables is 90%. Max scores tend to be larger when the two methods do not agree.

(PDF)

S4 Fig. Pathway scores for random phenotypes using max gene scores. P-values for 1077 pathways from our pathway library were computed for 100 random phenotypes using the *Pascal* pipeline using max gene scores and *chi-squared* pathway integration strategy (a) without merging of neighbouring genes and (b) with merging of neighbouring genes (gene-fusion strategy). P-value distributions are represented by QQ-plots (upper panels) and histograms (lower panels). Results are colour-coded according to the fraction of genes in a given pathway that have a neighbouring gene in the same pathway, i.e. that are located nearby on the genome (distance < 300kb). (a) P-values of pathways that contain genes in LD are strongly inflated without correction. (b) The gene fusion approach provides well-calibrated p-values independently of the number of pathway genes in LD.

(PDF)

S5 Fig. Performance of pathway enrichment methods for blood lipid traits and Crohn's disease using sum of squares (SOCS) statistics for defining gene scores. Displayed is the mean area under the precision-recall curve (AUC) for pathways identified using *Pascal*, a standard hypergeometric test at various gene score thresholds, and a rank-sum test (vertical bars show the standard error). We show results for the SOCS gene scores (MOCS gene score results are similar, see Fig 4 in the main text). a) Results for four blood lipid traits. A reference standard pathway list was defined as all pathways that show a significance level below 5×10^{-6} , for any of the tested threshold parameters for hypergeometric tests in the largest study of lipid traits to date. The significance level of 5×10^{-6} corresponds to the Bonferroni corrected, genome-wide significance threshold at the 0.5% level for a single method. For each phenotype, error bars denote the standard error computed from three independent subsamples of the *CoLaus* study (including 1500 individuals each). We see good overall performance of *Pascal* pathway scores, whereas results for discrete gene sets vary widely with the particular choice for the threshold

parameter of hypergeometric test. b) Results for Crohn's disease using the same approach as in (a). A reference standard pathway list was defined as all pathways that show a significance level below 5×10^{-6} for *any* of the tested threshold parameters for hypergeometric tests in the largest study of Crohn's disease traits to date. We observe that the chi-squared strategy outperforms all other strategies in this setting, whereas performance of the hypergeometric testing strategy varies.

(PDF)

S6 Fig. Power of pathway scoring methods across diverse traits and diseases using sum of squares (SOCS) statistics for defining gene scores. Bar heights represent the number of pathways found to be significant after Bonferroni correction. Within a given trait group, results are aggregated for all tested GWAS studies. 65 GWAS had at least one significant pathway in one of the tested method. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. This was done to avoid that a few studies with many emerging pathways dominate. We show results for the SOCS gene scores (MOCS gene score results are similar, see Fig 5). (a) Results are aggregated over all trait groups. (b) Results for different trait groups.

(PDF)

S7 Fig. Power of pathway scoring methods stratified with respect to sample size. Only GWAS studies for quantitative traits were used. Top panels (a,b) show results for max gene scores and bottom panels (c,d) show results for sum gene scores. (a,c) Results for all studies where the number of individuals was below 50,000. (b,d) Results for studies with sample sizes above 50,000. We see power gains in all cases. The improvements are particularly pronounced in lower powered GWAS.

(PDF)

S8 Fig. Power comparison max and sum gene scores for pathway analysis. Bar heights represent the number of pathways found to be significant after Bonferroni correction. Within a given trait group, results are aggregated for all tested GWAS studies. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. Results for SOCS and MOCS as well as the chi-square and empirical pathway scores are displayed. We observe a drop in performance for the combination of MOCS gene scores with empirical pathway scores.

(PDF)

S9 Fig. Power analysis for max gene scores with capped gene scores. Bar heights represent the number of pathways found to be significant after Bonferroni correction. Within a given trait group, results are aggregated for all tested GWAS studies. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. Max gene scores using empirical sampling pathway scores (emp) and chi-squared pathway scores (chi2) are compared to max gene scores combined with empirical sampling, where outlier gene scores ($p\text{-value} < 10^{-12}$) are set to 10^{-12} (empCapped). We chose the capping value such that the maximum $-\log_{10}$ p-value was roughly in the middle between genome wide significance threshold (8) and the maximum value that can be calculated for the sum statistic (15).

(PDF)

S10 Fig. Power of Pascal pathway scoring methods compared to aggregated hypergeometric scores (MOCS). The same data as in Fig 5 is plotted here. However, instead of comparing Pascal pathway scoring methods with results for all hypergeometric threshold separately, we

defined a new aggregated pathway score that picks the optimal threshold for each pathway over a range of hypergeometric threshold and correcting for the multiple number of tests by Bonferroni correction. Results for different sets of thresholds are displayed. Set1 refers to the complete set of thresholds (i.e.: 25%, 15%, 10%, 5%, 2%, 1%, 0.25%, 0.1%). Set2 refers to a set with thresholds more ‘spread out’ (i.e.: 25%, 5%, 1%, 0.25). We see that *Pascal* has better performance, except when combining the ‘empirical sampling’ pathway scoring method with max gene scores.

(PDF)

S11 Fig. Power of *Pascal* pathway scoring methods compared to aggregated hypergeometric scores (SOCS). The same data as in Fig 5 is plotted here. However, instead of comparing *Pascal* pathway scoring methods with results for all hypergeometric threshold separately, we defined a new aggregated pathway score that picks the optimal threshold for each pathway over a range of hypergeometric threshold and correcting for the multiple number of tests by Bonferroni correction. Results for different sets of thresholds are displayed. Set1 refers to the complete set of thresholds (i.e.: 25%, 15%, 10%, 5%, 2%, 1%, 0.25%, 0.1%). Set2 refers to a set with thresholds more ‘spread out’ (i.e.: 25%, 5%, 1%, 0.25). We see that *Pascal* has better performance.

(PDF)

S12 Fig. Power of gamma distribution for pathway analysis (MOCS). Bar heights represent the number of pathways found to be significant after Bonferroni correction. Different bars signify results for a different gamma shape parameter value. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. Upper left panel ‘All’ refers to all traits stacked. We present here MOCS gene score based results. 52 GWAS showed at least one significant pathway in one of the evaluated scenarios.

(PDF)

S13 Fig. Power of gamma distribution for pathway analysis (SOCS). Bar heights represent the number of pathways found to be significant after Bonferroni correction. Different bars signify results for a different gamma shape parameter value. For each GWAS, the raw number of significant pathways was divided by the number of pathways found by the best performing method. Upper left panel ‘All’ refers to all traits stacked. We present here MOCS gene score based results. 60 GWAS showed at least one significant pathway in one of the evaluated scenarios.

(PDF)

S14 Fig. Distribution of pathway scores for simulated phenotypes influenced by causal SNPs in coding regions. We first sampled 50 random SNPs assayed in *CoLaus* in or close to coding regions. Using the genotypes of the *CoLaus* study we then simulated phenotypes by adding up the sampled 50 SNPs with a normally distributed effect size with a variance of 0.04 plus Gaussian noise (with a variance of 1). We then ran GWAS for the simulated phenotype to obtain association summary statistics. The experiment was repeated 50 times. On average, this resulted in 18 independent, genome-wide significant gene score hits for each simulated GWAS (for the MOCS statistic). We applied *Pascal* to compute pathway scores for each of the 50 simulated GWAS. We found that the resulting pathway scores are well calibrated, i.e., they do not show inflation or deflation regardless of the setting used (max or sum gene score, chi2 or empirical enrichment test). The QQ-plots show the median value for each quantile across the 50 simulated GWAS. The shaded areas correspond to 95% confidence intervals for the median (estimated from 2000 bootstrap samples of size 50, with replacements). Similar results were

obtained when varying the type and number of simulated causal SNPs and their effect size. (PDF)

S15 Fig. Distribution of pathway scores for simulated phenotype influenced by causal SNPs in coding and non-coding regions. These QQ-plots correspond to an analysis equivalent to that of [S14 Fig](#) but with 50 SNPs chosen uniformly from all SNPs assayed in *CoLaus*, rather than from genic regions only. On average, this resulted in 12 independent, genome-wide significant gene score hits for each simulated GWAS (using the MOCS statistic). Note that this does not completely exclude the possibility of less well-calibrated scores in other settings. Deviations from perfectly calibrated scores may occur in the cases where true SNP associations are present, because the gene wise test statistic may have varying power for different genes depending on the genetic architecture of the associated phenotype and on certain gene properties (such as gene length, LD structure, SNP coverage, or SNP allele frequency). If a set of pathways contains many pathways enriched (or depleted) for genes with such confounding factors, inflation or deflation is possible. (PDF)

S1 Table. Details of GWAS used. Details of the 118 GWAS that we used for comparing *Pascal* with other methods. (TXT)

S2 Table. Pathways found by *Pascal*. Tables of pathways discovered by *Pascal* for the 118 GWAS. (TXT)

Author Contributions

Conceived and designed the experiments: DL DM ZK SB. Performed the experiments: DL DM. Analyzed the data: DL DM. Wrote the paper: DL DM RR ZK SB.

References

1. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet.* 2012; 90: 7–24. doi: [10.1016/j.ajhg.2011.11.029](https://doi.org/10.1016/j.ajhg.2011.11.029) PMID: [22243964](https://pubmed.ncbi.nlm.nih.gov/22243964/)
2. Hou L, Zhao H. A review of post-GWAS prioritization approaches. *Front Genet.* 2013; 4: 280. doi: [10.3389/fgene.2013.00280](https://doi.org/10.3389/fgene.2013.00280) PMID: [24367376](https://pubmed.ncbi.nlm.nih.gov/24367376/)
3. Segrè A V, Groop L, Mootha VK, Daly MJ, Altshuler D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemc traits. *PLoS Genet.* 2010; 6: e1001058. doi: [10.1371/journal.pgen.1001058](https://doi.org/10.1371/journal.pgen.1001058) PMID: [20714348](https://pubmed.ncbi.nlm.nih.gov/20714348/)
4. Pers TH, Karjalainen JM, Chan Y, Westra H-J, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun.* Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015; 6: 5890. Available: doi: <http://dx.doi.org/10.1038/ncomms6890> PMID: [25597830](https://pubmed.ncbi.nlm.nih.gov/25597830/)
5. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet.* 2014; 46: 1173–86. doi: [10.1038/ng.3097](https://doi.org/10.1038/ng.3097) PMID: [25282103](https://pubmed.ncbi.nlm.nih.gov/25282103/)
6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000; 25: 25–29. doi: [10.1038/75556](https://doi.org/10.1038/75556) PMID: [10802651](https://pubmed.ncbi.nlm.nih.gov/10802651/)
7. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res.* 2014; 42: D199–D205. doi: [10.1093/nar/gkt1076](https://doi.org/10.1093/nar/gkt1076) PMID: [24214961](https://pubmed.ncbi.nlm.nih.gov/24214961/)
8. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* 2003; 13: 2129–2141. doi: [10.1101/gr.772403](https://doi.org/10.1101/gr.772403) PMID: [12952881](https://pubmed.ncbi.nlm.nih.gov/12952881/)

9. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Res.* 2011; 39: D691–D697. doi: [10.1093/nar/gkq1018](https://doi.org/10.1093/nar/gkq1018) PMID: [21067998](https://pubmed.ncbi.nlm.nih.gov/21067998/)
10. Nishimura D. *BioCarta. Biotech Softw Internet Rep.* 2001; 2: 117–120. doi: [10.1089/152791601750294344](https://doi.org/10.1089/152791601750294344)
11. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet. The American Society of Human Genetics;* 2010; 87: 139–145. doi: [10.1016/j.ajhg.2010.06.009](https://doi.org/10.1016/j.ajhg.2010.06.009) PMID: [20598278](https://pubmed.ncbi.nlm.nih.gov/20598278/)
12. Li MX, Gui HS, Kwan JSH, Sham PC. GATES: A rapid and powerful gene-based association test using extended Simes procedure. *Am J Hum Genet.* 2011; 88: 283–293. doi: [10.1016/j.ajhg.2011.01.019](https://doi.org/10.1016/j.ajhg.2011.01.019) PMID: [21397060](https://pubmed.ncbi.nlm.nih.gov/21397060/)
13. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* 2011; 21: 1109–21. doi: [10.1101/gr.118992.110](https://doi.org/10.1101/gr.118992.110) PMID: [21536720](https://pubmed.ncbi.nlm.nih.gov/21536720/)
14. Wang L, Jia P, Wolfinger RD, Chen X, Grayson BL, Aune TM, et al. An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. *Bioinformatics.* 2011; 27: 686–692. doi: [10.1093/bioinformatics/btq728](https://doi.org/10.1093/bioinformatics/btq728) PMID: [21266443](https://pubmed.ncbi.nlm.nih.gov/21266443/)
15. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet.* 2007; 81: 1278–1283. doi: [10.1086/522374](https://doi.org/10.1086/522374) PMID: [17966091](https://pubmed.ncbi.nlm.nih.gov/17966091/)
16. Ehret GB, Lamparter D, Hoggart CJ, Whittaker JC, Beckmann JS, Kutalik Z. A multi-SNP locus-association method reveals a substantial fraction of the missing heritability. *Am J Hum Genet.* 2012; 91: 863–871. doi: [10.1016/j.ajhg.2012.09.013](https://doi.org/10.1016/j.ajhg.2012.09.013) PMID: [23122585](https://pubmed.ncbi.nlm.nih.gov/23122585/)
17. Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics.* 2012. pp. 369–375. doi: [10.1038/ng.2213](https://doi.org/10.1038/ng.2213) PMID: [22426310](https://pubmed.ncbi.nlm.nih.gov/22426310/)
18. Holmans P, Green EK, Pahwa JS, Ferreira M a R, Purcell SM, Sklar P, et al. Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder. *Am J Hum Genet.* 2009; 85: 13–24. doi: [10.1016/j.ajhg.2009.05.011](https://doi.org/10.1016/j.ajhg.2009.05.011) PMID: [19539887](https://pubmed.ncbi.nlm.nih.gov/19539887/)
19. Evangelou M, Smyth DJ, Fortune MD, Burren OS, Walker NM, Guo H, et al. A Method for Gene-Based Pathway Analysis Using Genomewide Association Study Summary Statistics Reveals Nine New Type 1 Diabetes Associations *Genetic Epidemiology. Genet Epidemiol.* 2014; 38: 661–670. doi: [10.1002/gepi.21853](https://doi.org/10.1002/gepi.21853) PMID: [25371288](https://pubmed.ncbi.nlm.nih.gov/25371288/)
20. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011; 89: 82–93. doi: [10.1016/j.ajhg.2011.05.029](https://doi.org/10.1016/j.ajhg.2011.05.029) PMID: [21737059](https://pubmed.ncbi.nlm.nih.gov/21737059/)
21. Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am J Hum Genet.* 2007; 81: 1158–1168. doi: [10.1086/522036](https://doi.org/10.1086/522036) PMID: [17966093](https://pubmed.ncbi.nlm.nih.gov/17966093/)
22. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012; 491: 56–65. doi: [10.1038/nature11632](https://doi.org/10.1038/nature11632) PMID: [23128226](https://pubmed.ncbi.nlm.nih.gov/23128226/)
23. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet.* 2013; 45: 1274–83. doi: [10.1038/ng.2797](https://doi.org/10.1038/ng.2797) PMID: [24097068](https://pubmed.ncbi.nlm.nih.gov/24097068/)
24. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature.* 2010; 467: 52–8. doi: [10.1038/nature09298](https://doi.org/10.1038/nature09298) PMID: [20811451](https://pubmed.ncbi.nlm.nih.gov/20811451/)
25. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature.* 2014; 506: 376–81. doi: [10.1038/nature12873](https://doi.org/10.1038/nature12873) PMID: [24390342](https://pubmed.ncbi.nlm.nih.gov/24390342/)
26. Mishra A, Macgregor S. VEGAS2 : Software for More Flexible Gene-Based Testing. *Twin Res Hum Genet.* 2015; 18: 86–91. doi: [10.1017/thg.2014.79](https://doi.org/10.1017/thg.2014.79) PMID: [25518859](https://pubmed.ncbi.nlm.nih.gov/25518859/)
27. Firmann M, Mayor V, Vidal P, Bochud M, Pécoud A, Hayoz D, et al. The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovascular Disorders.* 2008. p. 6. doi: [10.1186/1471-2261-8-6](https://doi.org/10.1186/1471-2261-8-6) PMID: [18366642](https://pubmed.ncbi.nlm.nih.gov/18366642/)
28. Heinig M, Petretto E, Wallace C, Bottolo L, Rotival M, Lu H, et al. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature.* 2010; 467: 460–464. doi: [10.1038/nature09386](https://doi.org/10.1038/nature09386) PMID: [20827270](https://pubmed.ncbi.nlm.nih.gov/20827270/)

29. Burren OS, Guo H, Wallace C. VSEAMS : A pipeline for variant set enrichment analysis using summary GWAS data identifies IKZF3, BATF and ESRRA as key transcription factors in type 1 diabetes. 2014;30: 0–26. doi: [10.1093/bioinformatics/btu571](https://doi.org/10.1093/bioinformatics/btu571)
30. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. Proc 23rd Int Conf Mach Learn—ICML'06. 2006; 233–240. doi: [10.1145/1143844.1143874](https://doi.org/10.1145/1143844.1143874)
31. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. Nature Publishing Group; 2010; 42: 1118–25. doi: [10.1038/ng.717](https://doi.org/10.1038/ng.717) PMID: [21102463](https://pubmed.ncbi.nlm.nih.gov/21102463/)
32. Imielinski M, Baldassano RN, Griffiths A, Russell RK, Annesse V, Dubinsky M, et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat Genet. 2009; 41: 1335–1340. doi: [10.1038/ng.489](https://doi.org/10.1038/ng.489) PMID: [19915574](https://pubmed.ncbi.nlm.nih.gov/19915574/)
33. Wellcome T, Case T, Consortium C. Genome-wide association study of 14, 000 cases of seven common diseases and. Nature. 2007; 447: 661–78. doi: [10.1038/nature05911](https://doi.org/10.1038/nature05911) PMID: [17554300](https://pubmed.ncbi.nlm.nih.gov/17554300/)
34. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet. 2010; 42: 105–116. doi: [10.1038/ng.520](https://doi.org/10.1038/ng.520) PMID: [20081858](https://pubmed.ncbi.nlm.nih.gov/20081858/)
35. Estrada K, Styrkarsdottir U, Evangelou E, Hsu YH, Duncan EL, Ntzani EE, et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. 2012;44: 491–501. doi: [10.1038/ng.2249](https://doi.org/10.1038/ng.2249)
36. Day TF, Yang Y. Wnt and hedgehog signaling pathways in bone development. J Bone Joint Surg Am. 2008; 90 Suppl 1: 19–24. doi: [10.2106/JBJS.G.01174](https://doi.org/10.2106/JBJS.G.01174) PMID: [18292352](https://pubmed.ncbi.nlm.nih.gov/18292352/)
37. Tobacco T, Consortium G. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet. 2010; 42: 441–7. doi: [10.1038/ng.571](https://doi.org/10.1038/ng.571) PMID: [20418890](https://pubmed.ncbi.nlm.nih.gov/20418890/)
38. Bradley DT, Zipfel PF, Hughes AE. Complement in age-related macular degeneration: a focus on function. Eye (Lond). 2011; 25: 683–693. doi: [10.1038/eye.2011.37](https://doi.org/10.1038/eye.2011.37)
39. Ebrahimi KB, Handa JT. Lipids, lipoproteins, and age-related macular degeneration. J Lipids. 2011; 2011: 802059. doi: [10.1155/2011/802059](https://doi.org/10.1155/2011/802059) PMID: [21822496](https://pubmed.ncbi.nlm.nih.gov/21822496/)
40. Lee D, Williamson VS, Bigdeli TB, Riley BP, Fanous a. H, Vladimirov VI, et al. JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. Bioinformatics. 2014; 31: 1176–1182. doi: [10.1093/bioinformatics/btu816](https://doi.org/10.1093/bioinformatics/btu816) PMID: [25505091](https://pubmed.ncbi.nlm.nih.gov/25505091/)
41. Genz A. Numerical Computation of Multivariate Normal Probabilities. J Comput Graph Stat. 1992; 1: 141–149. doi: [10.1080/10618600.1992.10477010](https://doi.org/10.1080/10618600.1992.10477010)
42. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 1999. pp. 29–34. doi: [10.1093/nar/27.1.29](https://doi.org/10.1093/nar/27.1.29) PMID: [9847135](https://pubmed.ncbi.nlm.nih.gov/9847135/)
43. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette M a, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102: 15545–50. doi: [10.1073/pnas.0506580102](https://doi.org/10.1073/pnas.0506580102) PMID: [16199517](https://pubmed.ncbi.nlm.nih.gov/16199517/)
44. Xu Z, Duan Q, Yan S, Chen W, Li M, Lange E, et al. DISSCO: direct imputation of summary statistics allowing covariates. Bioinformatics. 2015; 31: 2434–2442. doi: [10.1093/bioinformatics/btv168](https://doi.org/10.1093/bioinformatics/btv168) PMID: [25810429](https://pubmed.ncbi.nlm.nih.gov/25810429/)
45. Ehret GB, Lamparter D, Hoggart CJ, Whittaker JC, Beckmann JS, Kutalik Z. A multi-SNP locus-association method reveals a substantial fraction of the missing heritability. Am J Hum Genet. 2012; 91: 863–871. doi: [10.1016/j.ajhg.2012.09.013](https://doi.org/10.1016/j.ajhg.2012.09.013) PMID: [23122585](https://pubmed.ncbi.nlm.nih.gov/23122585/)
46. Gao X, Stamer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet Epidemiol. 2008; 32: 361–369. doi: [10.1002/gepi.20310](https://doi.org/10.1002/gepi.20310) PMID: [18271029](https://pubmed.ncbi.nlm.nih.gov/18271029/)
47. B DR. The Distribution of a Linear Combination of x^2 Random Variables. J R Stat Soc Ser C. 1980; 29: 323–333.
48. Farebrother R. Algorithm AS 204: the distribution of a positive linear combination of χ^2 random variables. J R Stat Soc Ser C. 1984; 33: 332–339. doi: [10.2307/2347721](https://doi.org/10.2307/2347721)
49. Duchesne P, Lafaye De Micheaux P. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. Comput Stat Data Anal. 2010; 54: 858–862. doi: [10.1016/j.csda.2009.11.025](https://doi.org/10.1016/j.csda.2009.11.025)