

# **FORGE: multivariate calculation of gene-wide p-values from Genome-Wide Association Studies**

## **Authors and Affiliations**

Inti Pedroso<sup>1,2</sup>, Michael R Barnes<sup>3</sup>, Anbarasu Lourdasamy<sup>2</sup>, Ammar Al-Chalabi<sup>4</sup>,  
Gerome Breen<sup>1,2\*</sup>.

<sup>1</sup> MRC SGDP Centre, Institute of Psychiatry, King's College London, De Crespigny Park, London, SE5 8AF, United Kingdom.

<sup>2</sup> NIHR Biomedical Research Centre for Mental Health, South London and Maudsley NHS Foundation Trust and Institute of Psychiatry, King's College London, London, De Crespigny Park, London, SE5 8AF, United Kingdom.

<sup>3</sup> Department of Computational Biology, GlaxoSmithKline, Gunnels Wood Road, Stevenage, Hertfordshire, United Kingdom.

<sup>4</sup> Medical Research Council Centre for Neurodegeneration Research, Department of Clinical Neuroscience, P043, Institute of Psychiatry, King's College London, London SE5 8AF, United Kingdom.

## Abstract

Genome-wide association studies (GWAS) have proven a valuable tool to explore the genetic basis of many traits. However, many GWAS lack statistical power and the commonly used single-point analysis method needs to be complemented to enhance power and interpretation. Multivariate region or gene-wide association are an alternative, allowing for identification of disease genes in a manner more robust to allelic heterogeneity. Gene-based association also facilitates systems biology analyses by generating a single p-value per gene. We have designed and implemented FORGE, a software suite which implements a range of methods for the combination of p-values for the individual genetic variants within a gene or genomic region. The software can be used with summary statistics (marker ids and p-values) and accepts as input the result file formats of commonly used genetic association software. When applied to a study of Crohn's disease susceptibility, it identified all genes found by single SNP analysis and additional genes identified by large independent meta-analysis. FORGE p-values on gene-set analyses highlighted association with the Jak-STAT and cytokine signalling pathways, both previously associated with CD. We highlight the software's main features, its future development directions and provide a comparison with alternative available software tools. FORGE can be freely accessed at <https://github.com/inti/FORGE>.

Contact: [gerome.breen@kcl.ac.uk](mailto:gerome.breen@kcl.ac.uk) and [intipedroso@gmail.com](mailto:intipedroso@gmail.com)

## Introduction

Genome-wide association studies (GWAS) have led to the discovery of hundreds of replicated variants associated with diverse human phenotypes <sup>1</sup>. However, most GWAS have low statistical power to detect true effects due to the low effect-sizes of common risk alleles and the need for robust genome-wide significance thresholds, to allow for multiple testing, e.g.  $7.2 \times 10^{-8}$  <sup>2</sup>. Multivariate analytical strategies, such as gene-wide association, are an attractive alternative to single SNP approaches. They have the capability to allow for allelic heterogeneity (independent associated alleles in the same region) (e.g. <sup>3</sup>), result in fewer tests genome-wide <sup>4</sup> and provide gene p-values that can be used with gene-set analysis methods <sup>5</sup>. Increasing evidence suggest that allelic heterogeneity is a common feature of the genetic architecture of complex traits and that gene-set methods can improve the interpretation and statistical power of GWAS by using prior biological knowledge <sup>6-8</sup>.

Numerous gene-wide association methods have been proposed, for example <sup>9-13,13,14</sup>. The simplest form consists of correcting the minimum p-value in the region or gene by its number of single nucleotide polymorphisms (SNPs), e.g., a Bonferroni correction or a measure of effective number of tests <sup>15,16</sup>. This approach ignores possible allelic heterogeneity, i.e., independent potentially associated alleles. Multivariate methods are an alternative but they can require computationally expensive simulations to derive significance if the test statistic's null distribution is unknown, as may be due to the correlation between genetic markers.

Currently available software to perform gene-based association include: i) VEGAS that implements a simulation-based strategy to estimate significance <sup>13</sup>; ii) PLINK

with an implementation of Hotelling's T2-statistics and Makambi's modified Fisher's test statistic<sup>17</sup>, --T2 and --set-screen options respectively; and iii) GATES that implements a modified Simes test<sup>14</sup>. PLINK also allows users to perform SNP-set analyses, e.g., all SNPs within genes of a biological pathway, allowing for gene-set analyses. Analysis of GWAS using gene-set analysis has been reviewed recently elsewhere, e.g.,<sup>5,6</sup>, and we provide a representative list of available tools and studies in Supplementary Table 1 and 2, respectively.

Here we describe FORGE, a software suit to perform gene-wide and gene-set analyses of GWAS. It provides routines for the calculation of four gene-wide association methods and two gene-set analysis strategies. FORGE represents an extension and complements parallel independent efforts, such as VEGAS<sup>13</sup> or KGG<sup>14,18</sup>. In addition, several utility programs are distributed with FORGE allowing users to: i) map SNP to Genes using the Ensembl human genome annotation; ii) parse different gene-set files; and iii) calculate meta-analysis statistics for gene and gene-sets analyses results when studies carried out on multiple data-sets.

## Methods

### Gene-wide statistics

#### 1. Sidak's correction on minimum p-value

Consider a set of  $m$  SNPs ( $\mathbf{M}$ ) and their association p-values ( $\mathbf{P}$ ) with a trait, for which the aim is to calculate a combined association test statistic for  $\mathbf{M}$ . The simplest strategy is to consider the minimum p-value among  $\mathbf{P}$  as  $\mathbf{M}$ 's evidence for association using Sidak's correction<sup>19</sup> with an estimate of the number of effective tests within the

gene,  $p_{sidak} = 1 - (1 - p_{raw})^k$ , where  $p_{raw}$  is the minimum p-value in **P** and  $k$  is the effective number of SNPs tested calculated with the method of <sup>16</sup>.

## 2. Modified Fisher's method to combine correlated p-values

In order to obtain a combined test (**T**) for **M** taking the correlation among the genetic markers into account, we use the method originally derived by Brown (1975) with the modifications proposed by Kost and McDermott <sup>20</sup> and Makambi <sup>17</sup>, leading to the chi-square test  $T = 0.5 \times M_{F,m}$ , with  $v$  degrees of freedom, where  $M_{F,m} = -2 \times \sum_{i=1}^m w_i \times \log(p_i)$  is the weighted version of the Fisher's method,  $p_i$  is the p-value of  $i$ th marker and  $w_i$  are weights greater than zero that sum to one. The degrees of freedom are  $v = 8 / \text{var}(M_{F,m})$  with  $\text{var}(M_{F,m}) = \sum_{i=1}^m \sum_{j=1}^m w_i w_j (3.263 |\rho_{ij}| + 0.71 |\rho_{ij}|^2 + 0.027 |\rho_{ij}|^3)$ , where  $\rho_{ij}$  is the correlation between the  $P_i$  and  $P_j$ .

## 3. Fixed-effect z-score statistic

We can also calculate  $Z_{fix} = (\sum_{i=1}^m z_i w_i / \sum_{i=1}^m w_i) \times \mathcal{N}^{0.5}$  (Huedo-Medina *et al.*, 2006), where  $z_i$  are the p-values transformed to z-scores using the standard normal distribution inverse cumulative distribution function (c.d.f.) and  $V_{fix}$  is the variance of the test. Using the approximation of the multivariate-normal distribution  $V_{fix} = \sum_{i=1}^m \sum_{j=1}^m w_i w_j \rho_{ij}^2$ .

### 1. Random-effect z-score statistic

A random-effects estimate is given by  $Z_{random} = (\sum_{i=1}^m z_i w_i^* / \sum_{i=1}^m w_i^*) \times \mathcal{N}^{0.5}$ , with variance  $V_{random} = \sum_{i=1}^m \sum_{j=1}^m w_i^* w_j^* \rho_{ij}^2$  and weights equal to  $w_i^* = (\tau^2 + w_i^{-1})^{-1}$  which are adjusted with the heterogeneity measure  $\tau^2 = \max[0, (Q - (m-1)) / (\sum w_i - \sum w_i^2 / \sum w_i)]$ . In calculating  $\tau^2$  one would normally use Cochran's heterogeneity statistics  $Q = \sum_{i=1}^m w_i (\sum_{i=1}^m z_i w_i / \sum_{i=1}^m w_i - z_i)^2$ , which is an approximately distributed chi-square variable with  $m-1$  degrees of freedom <sup>21</sup>. To account for the correlation among the genetic markers  $\tau^2$  is calculated using  $Q'$ , which is  $Q$  re-scaled into a chi-

square variable with  $m-1$  degrees of freedom. This is achieved by calculating  $Q$ 's tail probability using the modified Fisher's method described above and then  $Q$ ' is the probability's chi-square value from a chi-square distribution with  $m-1$  degrees of freedom.

## Gene-set analysis

### **1. SNP to gene-sets strategy**

In this case we treat gene-sets as a large gene, i.e. map to it all SNPs of its genes and applied the statistics described above.

### **2. Gene-sets analysis with gene p-values**

We implemented the methods described by Luo et al.<sup>22</sup> as following. GSA is performed by transforming the gene p-values into z-scores (using the standard distribution inverse c.d.f.) and combining the z-scores with  $S_{Net} = \left( \frac{\sum_{i=1}^k w_i g_i}{\sum_{i=1}^k w_i} \right) \sqrt{V_{SNet}}$  where  $g_i$  is the z-score of the  $i^{th}$  gene in the gene-set,  $k$  the number of genes in the gene-set and  $V_{SNet}$  is the variance-covariance matrix of the gene's statistic,

$$V_{SNet} = corr(g_i, g_j) = \frac{\sum_{i=1}^k \sum_{j=1}^k w_i w_j corr(z_{gi}, z_{gj})}{\sqrt{\left( \sum_{i=1}^k w_i \right) \left( \sum_{j=1}^k w_j \right)}} .$$

$S_{Net}$  is formally a variable from a standard normal distribution and its significance can be estimated with normal distribution probability density function.

## Calculation of gene p-values

The gene-wide statistics described above lead to asymptotic estimates of significance. We implemented the method described by Li et al.<sup>14</sup> to approximate the correlation between the p-values by the correlation between the SNPs, i.e. allelic LD or Pearson's correlation. In addition to these asymptotic strategy we implemented routines to calculate gene p-values using the simulation-based strategy of Liu et al.<sup>13</sup>. Although

this strategy is slower than asymptotic methods, its p-values are well correlated with empirical estimates<sup>13</sup>. We refer the reader to Liu et al.<sup>13</sup> for details of the strategy and to Supplementary Material for description of our implementation.

### *Analysis of the WTCCC Crohn's disease GWAS*

We obtained summary statistics of the Crohn's disease GWAS from the EGA website with formal data access permission of the WTCCC Data Access Committee. Quality control (QC) performed QC by excluding samples as indicated in the files provided by the WTCCC and SNPs with missingness  $\geq 1\%$ , minor allele frequency (MAF)  $\leq 1\%$  Hardy-Weinberg equilibrium (HWE) p-value  $\leq 1 \times 10^{-3}$  and p-value  $\leq 1 \times 10^{-5}$  in controls and cases, respectively, as previously described<sup>23</sup>. We also obtained genotype data of the bipolar disorder GWAS and performed QC as follow: i) poor quality samples and SNPs were removed as indicated in the files distributed by the WTCCC; and ii) SNPs were removed using the same criteria used for the Crohn's disease summary statistics. SNP association were performed with a logistic regression as implemented on PLINK<sup>24</sup>. Using both sets of summary statistics gene-wide statistics were calculated for approximately 19,550 protein-coding, long intergenic non-coding RNA and micro-RNA genes annotated in Ensembl version 59 and whose SNPs passed QC in the WTCCC studies. We mapped SNPs to genes if the SNP was within 20 kb of the annotated coordinates aiming to include 95% of potential eQTL loci<sup>25</sup>. We approximated the correlation between test statistics as the correlation between the SNPs as measured by the Person's correlation between the allele counts. We used a False Discovery Rate (FDR)  $< 0.1$ <sup>26</sup> to perform multiple testing correction on the gene p-values.

We performed gene-set association with the WTCCC Crohn's disease gene-wide p-values results using 5,384 gene-sets derived from the Human Protein Reference Database protein-protein interaction network (PPIN)<sup>27</sup>. The PPIN gene-sets were constructed with the following algorithm: subnetwork searches started from each node (seed node) in the PPIN and a subnetwork was defined by adding sequentially the direct neighbours of the subnetwork's nodes (initially only the seed node). We allowed searches to go to a max of 5 interactions from the seed node and generate subnetworks of 2 to 200 nodes in size; each subnetwork was used as a gene-set. An FDR of 0.1 was used for multiple testing correction<sup>28</sup>. Significant PPIN's genes were analysed by assessing their over-representation among biological categories reported on KEGG<sup>29</sup> and GO databases<sup>30</sup>. Significance of the overlap was calculated with binomial statistics.

## Results and Discussion

Figure 1 presents the GWAS analyses implemented in FORGE. GWAS summary statistics are used to calculate gene-set p-values either by mapping SNP p-values to gene-sets directly or by calculating gene-based p-values as an intermediate step. Calculation of gene p-values provides an additional layer of analysis that can facilitate integration of GWAS with results from other “omics” technologies that also generate a single statistic per gene, e.g., gene-expression studies. This integration can also be performed at the level of gene-sets.

Liu et al.<sup>13</sup> introduced a simulation-based gene-wide association strategy that provides gene p-values with very good agreement with those obtained by phenotype permutations. For all gene-wide association methods we implemented this strategy as a way to estimate significance. This approach requires computation of correlations

between SNP genotypes. It is desirable to use genotypes from a reference population, e.g. HapMap project samples, because it enables the use of summary statistics in the absence of each study's genotype data. Using the WTCCC bipolar disorder GWAS we calculated gene p-values using the WTCCC genotypes and the HapMap Phase 3 CEU samples. The computing time decreased approximately linearly with the sample size, i.e., the computing time is reduced by ~50 times when using the 165 HapMap Phase 3 CEU samples compared with using the full WTCCC Crohn's disease dataset (5000 samples) (not shown). Also there was high correlation ( $>0.99$ ) between the gene p-values calculated with both sets of genotypes (not shown), as may have been expected due to the Caucasian origin of both samples. There was high correlation between all gene-wide p-values (minimum correlation  $> 0.7$ ) and the correlations between multivariate methods were higher (minimum correlation  $> 0.95$ ) (Figure 2). Importantly, the distribution of gene p-values for the bipolar disorder GWAS did not present over-dispersion (Figure 3), in agreement with the idea that this particular GWAS was underpowered to detect the true genetic effects<sup>31</sup>. Finally, we compared the results of the fixed and random-effects models and found small differences but only on relatively small genes, i.e., approximately  $< 45$  SNPs (Figure 4). For larger genes the statistical heterogeneity measure  $I^2$  is rarely different from zero. This pattern is to be expected since only a small fraction of genetic variation is thought to be associated with a phenotype, so in larger genes most of the evidence will point to lack of association and statistics will be more homogeneous (low  $I^2$ ). Thus, our exploration of using a statistical heterogeneity measure to tackle genetic heterogeneity suggest it is hardly effective in moderate to large genes. Application of other statistical techniques may provide better results, for example<sup>32,33</sup>.

### Comparison with other software

Table 1 presents a comparison of FORGE with other software available to perform gene and gene-set analyses on GWAS. Compared with most other software, FORGE allows to perform both gene and gene-set analyses functionalities within the same software. In addition, it implements more analysis methods including asymptotic and simulation-based calculations. FORGE also reads all three major genotype formats used in GWAS and provides utilities to build SNP-to-gene mapping with updated versions of the human genome annotation. Finally, its Perl implementation makes it platform independent.

### Case study: the WTCCC Crohn's disease GWAS

Wang et al.<sup>6</sup> recently reviewed gene-set analyses results of GWAS of CD and we use these as a reference for compare results. FORGE gene-based association was performed using summary statistics and genotypes of the HapMap phase 3 CEU samples genotypes. Table 2 lists genes with  $FDR < 0.1$  on the Z FIX gene-based association method. Genes mapped to significant SNPs ( $p\text{-values} < 5 \times 10^{-8}$ ) were also identified by the gene-based association. Interestingly, in line with results of Liu et al.<sup>13</sup> and Li et al.<sup>14</sup>, gene-based association results highlighted genes associated at a genome-wide significant level in other studies, in spite of having sub-threshold association on the GWAS analysed here. We used gene p-values calculated with the Z FIX method for gene set analyses (Table 3). Several significant associations were found among biological process with previous compelling evidence of involvement of Crohn's disease pathology<sup>6</sup>, for example: associations with gene-sets of the Jak-STAT (hsa04630) and Cytokine-cytokine receptor interaction (hsa04060) signalling pathways.

Overall, FORGE produces gene p-values that complement single SNP associations. Its gene p-values can be calculated from summary statistics and used with gene-set analysis methods to provide a systems biology perspective of a GWAS. Together, gene and gene-set association represent a complement to single SNP analyses by helping to interpret and extract information from GWAS.

## **Availability and Future Directions**

FORGE has been deposited in the public repository GitHub (<https://github.com>). Software code is updated using control version with GIT (<http://git-scm.com>) and users can access the latest stable or development versions at <https://github.com/inti/FORGE>. Current development is focused on a web-server version of FORGE to run on the 264 CPU computer cluster hosted at the NIHR BRC Centre for Mental Health (Institute of Psychiatry, KCL, UK). This interface will allow users to upload GWAS summary statistics and perform both gene and gene-set analysis online.

## **Design and Implementation**

We implemented these methods in a software suite called FORGE written in Perl ([www.perl.com](http://www.perl.com)) using the PDL, PDL::Stats and PDL::LinearAlgebra libraries, all freely available at the Comprehensive Perl Archive Network ([www.cpan.org](http://www.cpan.org)). These libraries allow performing calculation on double mathematical precision with the General Scientific Library (GSL) and the LAPACK library. All major functions have been implemented as separate libraries, for example: i) GWAS\_STATS.pm with statistical analysis routines; ii) GWAS\_IO.pm to deal with file formats commonly used in GWAS studies; and iii) CovMatrix.pm implements method to calculate correlations and covariance matrices. This design allows new features to be developed

in a modular fashion and the use of specific functions (e.g., reading binary genotype files) from software by the wider scientific community.

Three major programs were developed: i) `forge.pl` to perform gene and gene-set analyses; ii) `gsa.pl` implements additional routines to perform a gene-set analyses; and iii) `meta_analysis.pl` implements to combined results from independent studies.

## **Main features**

1. Input files: the program reads input and output files and genotype file formats of commonly used GWAS analysis software, e.g., genotype files in Pedigree, Pedigree Binary Format and SNP association files produced by PLINK <sup>24</sup>. Gene-set definitions are accepted in GMT format, described elsewhere ([www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Main\\_Page](http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Main_Page)).
2. SNP-SNP correlations: three measures of SNP-SNP correlations are implemented: i) shrinkage correlation estimate described by Schafer and Strimmer <sup>34</sup>; ii) LD; and iii) correlation among the test statistics as explained by Li et al. <sup>14</sup>.
3. SNP-to-Gene annotations: pre-computed files with genetic variants mapped to genes up to 500 kb from gene coordinates are available at the FORGE website. A Perl script to update the annotation using the Ensembl API <sup>35</sup> is distributed as a utility of FORGE.
4. Additional features: i) user provided SNP weights, e.g. functional scores, can be used and will be re-scaled to sum to 1 within each gene; ii) genomic-control correction <sup>36</sup> can be automatically performed within the program; iii) analyses can be restricted to chromosomes, gene lists or gene types (e.g. protein coding or miRNA genes); iv) Affymetrix SNP identifiers are accepted and mapped to rsids internally; and vi) gene-sets for major databases like KEGG <sup>29</sup> or GO <sup>30</sup> are

provided as well as those derived from the protein-protein interaction network (see Methods).

5. Documentation: Example files and tutorials are available on the software's website.

## Acknowledgments

Funding: IP, GB and AAC thank funding from the NIHR Biomedical Research Centre for Mental Health at the South London and Maudsley NHS Foundation Trust and Institute of Psychiatry, Kings College London. GB and AAC thank the Medical Research Council. GB also thanks the Wellcome Trust, NARSAD and BBSRC for funding. AAC thanks The Motor Neurone Disease Association of Great Britain and Northern Ireland, ALS Therapy Alliance, Angel Fund and ALS Association for support.

Conflict of Interest: MB is a full time employee and stock of GlaxoSmithKline. All other authors declare no conflict of interest.

## References

1. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U S A* **106**, 9362 (2009).
2. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227-234 (2008).

3. Furney, S. J. et al. Genome-wide association with MRI atrophy measures as a quantitative trait locus for Alzheimer's disease. *Mol. Psychiatry*. **16**, 1130-1138 (2011).
4. Neale, B. M. & Sham, P. C. The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet.* **75**, 353-362 (2004).
5. Pedroso, I. & Breen, G. in *Genetics of Complex Human Diseases* (eds Al-Chalabi, A. & Almasy, L.) 194-204 (Cold Spring Harbor Laboratory Press, NY, 2009).
6. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* **11**, 843-854 (2010).
7. Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-838 (2010).
8. Ideker, T., Dutkowski, J. & Hood, L. Boosting Signal-to-Noise in Complex Biology: Prior Knowledge Is Power. *Cell* **144**, 860-863 (2011).
9. Wessel, J. & Schork, N. J. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am. J. Hum. Genet.* **79**, 792-806 (2006).
10. Chapman, J. & Whittaker, J. Analysis of multiple SNPs in a candidate gene or region. *Genet. Epidemiol.* **32**, 560-566 (2008).
11. Moskvina, V. et al. Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. *Mol. Psychiatry*. **14**, 252-260 (2008).
12. Schaid, D. J. Evaluating associations of haplotypes with traits. *Genet. Epidemiol.* **27**, 348-364 (2004).

13. Liu, J. Z. et al. A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87**, 139-145 (2010).
14. Li, M. X., Gui, H. S., Kwan, J. S. & Sham, P. C. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am. J. Hum. Genet.* **88**, 283-293 (2011).
15. Galwey, N. W. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genet. Epidemiol.* **33**, 559-568 (2009).
16. Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol.* **32**, 361-369 (2008).
17. Makambi, K. Weighted inverse chi-square method for correlated significance tests. *J. Appl. Stat.* **30**, 225-234 (2003).
18. Li, M. X., Sham, P. C., Cherny, S. S. & Song, Y. Q. A knowledge-based weighting framework to boost the power of genome-wide association studies. *PLoS ONE* **5**, e14480 (2010).
19. Sidak, Z. On probabilities of rectangles in multivariate Student distributions: their dependence on correlations. *Ann. Math. Statist* **42**, 169-175 (1971).
20. Kost, J. T. & McDermott, M. P. Combining dependent p-values. *Stat. Probab. Lett.* **60**, 183-190 (2002).
21. Huedo-Medina, T. B., Sanchez-Meca, J., Marin-Martinez, F. & Botella, J. Assessing heterogeneity in meta-analysis: Q statistic or I<sup>2</sup> index? *Psychol Methods* **11**, 193-206 (2006).

22. Luo, L. et al. Genome-wide gene and pathway analysis. *Eur J Hum Genet* **18**, 1045-1053 (2010).
23. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).
24. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007).
25. Veyrieras, J. B. et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4**, e1000214 (2008).
26. Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* **96**, 1151-1160 (2001).
27. Keshava Prasad, T. S. et al. Human Protein Reference Database--2009 update. *Nucleic Acids Res.* **37**, D767-72 (2009).
28. Strimmer, K. A unified approach to false discovery rate estimation. *BMC bioinformatics* **9**, 303 (2008).
29. Kanehisa, M. et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480-4 (2008).
30. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29 (2000).
31. Craddock, N. & Sklar, P. Genetics of bipolar disorder: successful start to a long journey. *Trends Genet.* **25**, 99-105 (2009).
32. Wang, L. et al. An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. *Bioinformatics* **27**, 686-692 (2011).

33. Lebrech, J. J., Stijnen, T. & van Houwelingen, H. C. Dealing with heterogeneity between cohorts in genomewide SNP association studies. *Stat Appl Genet Mol Biol* **9**, Article 8 (2010).
34. Schafer, J. & Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol* **4**, Article32 (2005).
35. Rios, D. et al. A database and API for variation, dense genotyping and resequencing data. *BMC Bioinformatics* **11**, 238 (2010).
36. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).
37. Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118-1125 (2010).

## Figure Legends

**Figure 1. GWAS analyses implemented on FORGE.** A FORGE analysis starts with GWAS summary statistics. It is possible to calculate gene-set statistics directly from SNP association p-values or by calculating gene p-values as an intermediate step. In all steps FORGE introduces corrections for the LD to avoid inflation of the statistics. Yellow boxes present the advantages and potential of each GWAS result level.

**Figure 2. Correlation between gene-based p-values.** Black diagonal line represents the 1-to-1 correlation. Blue line is a trend calculated with a linear model. Correlation estimates between the methods are indicated on the table.

**Figure 3. Quantile-quantile plots of simulation-based gene p-values.** Plotted is the expected (x-axis) against the observed (y-axis)  $-\log_{10}$  of the gene p-values. Dotted line marks the 95 % confidence interval.

**Figure 4. Comparison of fix and random-effects estimates.** Plotted for each genes (N ~20000) is the  $\log_{10}$  of the ratio of fix and random-effects gene p-values. Insert (top right corner) presents data for the complete range of number of effective tests. Points are coloured (see figure legend) as the gene statistical heterogeneity measure  $I^2$ , which has range from 0 (no evidence of heterogeneity) to 100% (maximum evidence of heterogeneity).

## Tables

<i>Software</i>	<i>VEGAS</i>	<i>PLINK</i>	<i>ALIGATOR</i>	<i>dmGWAS</i>	<i>GenGen</i>	<i>GATES*</i>	<i>FORGE</i>
<b>Analysis type</b>							
Gene-based association	X	X				X	X
Gene-sets		X	X	X	X	X	X
<b>Statistics used</b>							
Best SNP	X				X	X	X
Fix-effects	X	X	X	X	X	X	X
Random-effects							
Correction for LD	X		X			X	X
<b>P-value calculation</b>							
Asymptotic method		X				X	X
Simulation-based significance	X		X				X
Use of SNP weights						X	X
Phenotype Permutations		X		X	X		
Gene shuffling (only for gene-sets)					X		X
<b>Input formats</b>							
Pedigree			a	a	a		X
Binary pedigree	X	X	a	a	a	X	X
Genotype probabilities		X	a	a	a		X
<b>SNP to gene mapping files</b>							
Available	X	X	X	X	X	X	X
Update/custom build						X	X
Summary statistics	X		X	X	X	X	X

**Table 1. Comparison of FORGE with other software to perform gene and gene-set analyses of GWAS.**<sup>a</sup> = method used by software does not need genotype files. \* reported features correspond to the KGG software (<http://bioinfo1.hku.hk:13080/kggweb/>) where GATES has been made available.

Gene Symbol	Position	Uncorrected Minimum p-value	Z FIX	FDR
CYLD <sup>a</sup>	17q12	$4 \times 10^{-14}$	$< 1 \times 10^{-6}$	$1 \times 10^{-3}$
ATG16L1 <sup>a</sup>	5q31.1	$4 \times 10^{-13}$	$< 1 \times 10^{-6}$	$1 \times 10^{-3}$
NKX2-3 <sup>a</sup>	16q12.1	$1 \times 10^{-8}$	$< 1 \times 10^{-6}$	$1 \times 10^{-3}$
IL23R <sup>a</sup>	2q37.1	$3 \times 10^{-12}$	$< 1 \times 10^{-6}$	$1 \times 10^{-3}$
NOD2 <sup>a</sup>	3p21.31	$3 \times 10^{-14}$	$< 1 \times 10^{-6}$	$1 \times 10^{-3}$
C5orf56 <sup>a</sup>	2p13.1	$1 \times 10^{-6}$	$< 1 \times 10^{-6}$	$1 \times 10^{-3}$
IRGM <sup>a</sup>	2p13.1	$1 \times 10^{-7}$	$< 1 \times 10^{-6}$	$1 \times 10^{-3}$
ZNF300 <sup>a</sup>	1q32.1	$1 \times 10^{-7}$	$2 \times 10^{-6}$	0.02
PTPN2 <sup>a</sup>	10q24.2	$1 \times 10^{-7}$	$2 \times 10^{-6}$	0.02
MST1 <sup>a</sup>	5q31.1	$1 \times 10^{-6}$	$3 \times 10^{-6}$	0.02
ENSG00000249738	5q33.3	$5 \times 10^{-5}$	$1 \times 10^{-5}$	0.09
ENSG00000221733 <sup>a</sup>	2p13.1	$2 \times 10^{-11}$	$2 \times 10^{-5}$	0.09
APEH <sup>a</sup>	5q33.1	$1 \times 10^{-6}$	$4 \times 10^{-6}$	0.09
HLA-DQA2	2p13.1	$5 \times 10^{-5}$	$3 \times 10^{-5}$	0.09
KIF21B <sup>a</sup>	19q13.31	$7 \times 10^{-5}$	$3 \times 10^{-5}$	0.09
CCL18	1p31.3	$1 \times 10^{-4}$	$5 \times 10^{-5}$	0.09
TAP2	3p21.31	$5 \times 10^{-5}$	$5 \times 10^{-5}$	0.09
HLA-DOB	3p21.31	$5 \times 10^{-5}$	$5 \times 10^{-5}$	0.09
ENSG00000250264	16q12.1	$5 \times 10^{-5}$	$5 \times 10^{-5}$	0.09
RNF123 <sup>a</sup>	17q21.2	$4 \times 10^{-5}$	$5 \times 10^{-5}$	0.09
ZNF283	8p23.1	$1 \times 10^{-4}$	$6 \times 10^{-5}$	0.09
LOXL3	3p21.31	$2 \times 10^{-4}$	$6 \times 10^{-5}$	0.09
DAG1	3p21.31	$3 \times 10^{-5}$	$7 \times 10^{-5}$	0.09
IRF1	3p21.31	$4 \times 10^{-6}$	$7 \times 10^{-5}$	0.09
P4HA2	18p11.21	$4 \times 10^{-4}$	$8 \times 10^{-5}$	0.09
STAT3 <sup>a</sup>	5q31.1	$2 \times 10^{-5}$	$8 \times 10^{-5}$	0.09
USP4	6p21.32	$6 \times 10^{-5}$	$8 \times 10^{-5}$	0.09
ENSG00000245942	6p21.32	$3 \times 10^{-4}$	$8 \times 10^{-5}$	0.09
GMPPB	1p31.3	$1 \times 10^{-4}$	$9 \times 10^{-5}$	0.09
C2orf65	6p21.32	$2 \times 10^{-4}$	$9 \times 10^{-5}$	0.09
TNKS	5q33.1	$1 \times 10^{-4}$	$9 \times 10^{-5}$	0.09
HTRA2	6p21.32	$2 \times 10^{-4}$	$1 \times 10^{-4}$	0.09

**Table 2. Genes with FDR < 0.1 from WTCCC CD GWAS.** Z FIX p-value was calculated with the simulation-based strategy using up to  $10^6$  simulations. FDR = false discovery rate. <sup>a</sup> Genes within 20 kb of genome-wide significant SNPs from the CD GWAS meta-analysis reported by Franke et al. <sup>37</sup>.

	Enrichment p-value	FDR	Biological Categories
PPIN-4224	$8 \times 10^{-11}$	0.01	hsa04630: Jak-STAT signaling pathway; hsa04060: Cytokine-cytokine receptor interaction
PPIN-3507	$2 \times 10^{-6}$	0.07	hsa04012: ErbB signaling pathway; hsa04910: Insulin signaling pathway
PPIN-2218	$4 \times 10^{-5}$	0.07	GO0008236: serine-type peptidase activity
PPIN-6093	$4 \times 10^{-5}$	0.07	hsa04010: MAPK signaling pathway
PPIN-1129	$4 \times 10^{-5}$	0.07	hsa04664: Fc epsilon RI signaling pathway; hsa04012: ErbB signaling pathway
PPIN-6841	$6 \times 10^{-5}$	0.07	hsa04020: Calcium signaling pathway
PPIN-4434	$6 \times 10^{-5}$	0.07	GO0006888: ER to Golgi vesicle-mediated transport; GO0045576: mast cell activation

**Table 3. Protein-protein interaction networks with FDR < 0.1 on CD GWAS.**  
Reported biological categories were over-represented among the subnetwork genes.

## Figures







