

GUESS-ing Polygenic Associations with Multiple Phenotypes Using a GPU-Based Evolutionary Stochastic Search Algorithm

Leonardo Bottolo^{1,9*}, Marc Chadeau-Hyam^{2,9}, David I. Hastie², Tanja Zeller³, Benoit Liquet^{4,5}, Paul Newcombe⁵, Loic Yengo^{6,7}, Philipp S. Wild⁸, Arne Schillert⁹, Andreas Ziegler⁹, Sune F. Nielsen^{10,11}, Adam S. Butterworth¹², Weang Kee Ho¹², Raphaële Castagné¹³, Thomas Munzel¹⁴, David Tregouet¹², Mario Falchi¹⁵, François Cambien¹³, Børge G. Nordestgaard^{10,11}, Frédéric Fumeron^{16,17}, Anne Tybjærg-Hansen¹¹, Philippe Froguel^{6,7,15,18}, John Danesh¹², Enrico Petretto¹⁹, Stefan Blankenberg³, Laurence Tiret¹³, Sylvia Richardson^{5*}

1 Department of Mathematics, Imperial College London, London, United Kingdom, **2** Department of Epidemiology and Biostatistics, Imperial College London, London, United Kingdom, **3** University Heart Center Hamburg, Department of General and Interventional Cardiology, Hamburg, Germany, **4** INSERM U897, University Victor Segalen, Bordeaux, France, **5** MRC Biostatistics Unit, Institute of Public Health, Cambridge, United Kingdom, **6** European Genomic Institute for Diabetes, Lille, France, **7** CNRS UMR 8199 - Institut Pasteur de Lille, Lille, France, **8** Clinical Epidemiology, Center for Thrombosis and Haemostasis, University Medical Center Mainz, Mainz, Germany, **9** Institute of Medical Biometry and Statistics, University of Lübeck, Lübeck, Germany, **10** Department of Clinical Biochemistry, Herlev Hospital, Copenhagen, Denmark, **11** Copenhagen University Hospital, University of Copenhagen, Copenhagen, Denmark, **12** Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom, **13** INSERM UMRS 937, Pierre and Marie Curie University (UPMC, Paris 6), Paris, France, **14** Department of Medicine II, University Medical Center Mainz, Mainz, Germany, **15** Department of Genomics of Common Disease, School Of Public Health, Hammersmith Hospital, Imperial College London, London, United Kingdom, **16** INSERM U695, Paris, France, **17** Université Paris Diderot-Paris 7, UFR de Médecine Site Bichat, Paris, France, **18** Université de Lille 2, Lille, France, **19** Medical Research Council Clinical Sciences Centre, Faculty of Medicine, Imperial College London, London, United Kingdom

Abstract

Genome-wide association studies (GWAS) yielded significant advances in defining the genetic architecture of complex traits and disease. Still, a major hurdle of GWAS is narrowing down multiple genetic associations to a few causal variants for functional studies. This becomes critical in multi-phenotype GWAS where detection and interpretability of complex SNP(s)-trait(s) associations are complicated by complex Linkage Disequilibrium patterns between SNPs and correlation between traits. Here we propose a computationally efficient algorithm (GUESS) to explore complex genetic-association models and maximize genetic variant detection. We integrated our algorithm with a new Bayesian strategy for multi-phenotype analysis to identify the specific contribution of each SNP to different trait combinations and study genetic regulation of lipid metabolism in the Gutenberg Health Study (GHS). Despite the relatively small size of GHS ($n = 3,175$), when compared with the largest published meta-GWAS ($n > 100,000$), GUESS recovered most of the major associations and was better at refining multi-trait associations than alternative methods. Amongst the new findings provided by GUESS, we revealed a strong association of *SORT1* with TG-APOB and *LIPC* with TG-HDL phenotypic groups, which were overlooked in the larger meta-GWAS and not revealed by competing approaches, associations that we replicated in two independent cohorts. Moreover, we demonstrated the increased power of GUESS over alternative multi-phenotype approaches, both Bayesian and non-Bayesian, in a simulation study that mimics real-case scenarios. We showed that our parallel implementation based on Graphics Processing Units outperforms alternative multi-phenotype methods. Beyond multivariate modelling of multi-phenotypes, our Bayesian model employs a flexible hierarchical prior structure for genetic effects that adapts to any correlation structure of the predictors and increases the power to identify associated variants. This provides a powerful tool for the analysis of diverse genomic features, for instance including gene expression and exome sequencing data, where complex dependencies are present in the predictor space.

Citation: Bottolo L, Chadeau-Hyam M, Hastie DI, Zeller T, Liquet B, et al. (2013) GUESS-ing Polygenic Associations with Multiple Phenotypes Using a GPU-Based Evolutionary Stochastic Search Algorithm. *PLoS Genet* 9(8): e1003657. doi:10.1371/journal.pgen.1003657

Editor: Greg Gibson, Georgia Institute of Technology, United States of America

Received: December 4, 2012; **Accepted:** May 30, 2013; **Published:** August 8, 2013

Copyright: © 2013 Bottolo et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Leonardo Bottolo received funding from the Wellcome Trust Value-in-People award. Leonardo Bottolo and Sylvia Richardson acknowledge support from the Medical Research Council (grant G1002319). Marc Chadeau-Hyam acknowledges support from European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 226756 (Envirogenomarkers to SA Kyrtopoulos). Enrico Petretto acknowledges funding from European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. HEALTH-F4-2010-241504 (EURATRANS) and Medical Research Council UK. David Tregouet and Enrico Petretto acknowledge support from PHC ALLIANCE 2009 (19419PH) grant. The Gutenberg Health Study is funded through the government of Rheinland-Pfalz ("Stiftung Rheinland Pfalz für Innovation", contract number AZ 961-386261/733), the research programs "Wissen schafft Zukunft" and "Schwerpunkt Vaskuläre Prävention" of the Johannes Gutenberg-University of Mainz and its contract with Boehringer Ingelheim and PHILIPS Medical Systems including an unrestricted grant for the Gutenberg Health Study. The research reported in this article was supported by the National Genome Network "NGFNplus" by the Federal Ministry of Education and Research, Germany (contract number project A3 01GS0833) and by a joint funding from the Federal Ministry of Education and Research, Germany (contract BMBF 01KU0908A) and from the Agence Nationale de la Recherche, France (contract ANR 09 GENO 106 01) for the project CARDomics. 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.

* E-mail: l.bottolo@imperial.ac.uk (LB); sylvia@mrc-bsu.cam.ac.uk (SR)

† These authors contributed equally to this work.

Introduction

This paper builds upon recent developments in Bayesian Variable Selection (BVS) to propose a novel strategy for studying the association between large sets of predictors (SNP, copy number variants, exome sequencing variants, gene expression and protein levels) and groups of correlated traits (i.e., outcomes). Such data commonly arise in Genome-Wide Association Studies (GWAS), where a large range of continuous phenotypes are recorded together with hundreds of thousands genetic markers [1], [2] as well as more widely in integrative genomics analyses. Our strategy is formulated within the linear model, a framework suited to the analysis of multiple continuous responses, and enhanced with a powerful stochastic search engine that explores the vast set of possible multivariate SNPs models, i.e. models involving different linear combinations of subsets of covariates. We take advantage of the existing Bayesian framework for multiple outcomes [3], [4], [5] and employ a conjugate hierarchical prior setup for genetic effects that adapts to any correlation structure among the predictors [6], [7], integrating over model uncertainty. The resulting model and associated novel GUESS (Graphical Unit Evolutionary Stochastic Search) implementation, enables the search for sparse sets of explanatory features at the genome-wide scale that are simultaneously associated with a set of continuous responses. We provide synthetic measures of evidence both for multivariate predictive models and for the marginal associations with each group of phenotypes, through the computation of the Model Posterior Probabilities (MPP), Marginal Posterior Probabilities of Inclusion (MPPI) and Bayes Factors (BFs).

Our strategy exploits the advantages provided by two approaches used in genetic association studies: firstly, the use of BVS to go beyond "single SNP analyses" in GWAS [8]; secondly, the joint modelling of multiple traits. This yields increased power and enhanced interpretability of the genetic associations, providing new insights into the underlying regulatory mechanisms. To the best of our knowledge, GUESS is the first integrated Bayesian computational tool that is able to perform both fast and efficient variable selection in large dimensional covariate space and association analyses with multiple continuous phenotypes. In a real case study of several blood lipid traits, we compared GUESS with two recently proposed Bayesian alternatives, namely the piMASS algorithm [8] and the Bayesian method that is included in the SNPTEST software [9]. In a simulation study that mimics real-case scenarios, we also compared GUESS with well-established non-Bayesian multi-phenotype approaches, namely Multivariate ANOVA [10], Multiple Group LASSO [11] and Sparse PLS [12]. Alternative machine learning strategies for GWAS [13] that require filtering the genetic markers in a pre-processing step or use "evolutionary computation" to detect the

best combination of genetic markers that predict the variation of the traits are not yet tailored to analyze multiple traits.

Advantages over alternative GWAS Bayesian methods

The recently proposed piMASS algorithm implements a BVS strategy for genome-wide association analysis of single continuous phenotypes with a novel prior specification for the variance of the regression coefficients. The implementation of piMASS is based on a single chain Monte Carlo Markov Chain (MCMC) algorithm tuned to analyse a single response, with the aim of demonstrating the feasibility of BVS in a model space with many predictors whilst showing the benefits of considering multivariate SNPs models and model uncertainty. However the specific proposal density used in the MCMC and implemented in piMASS cannot be extended easily in a multi-phenotype setup.

Our algorithm, GUESS, also considers BVS for such a large model space through an Evolutionary Stochastic Search algorithm [7], but differs from piMASS in three main aspects. Firstly, it is adapted to analyse either single or multiple phenotypes. Secondly, GUESS adopts sparsity-induced prior specification that helps the search algorithm to focus on models that are well supported by the data [8], allowing the user to specify natural quantities such as the prior expectation and standard deviation of the number of associated features. Lastly, GUESS uses an advanced stochastic search MCMC algorithm that is specifically designed to deal with the multi-modality of the model space [7], [14], [15], which potentially can contain competing sets of explanatory variables. The latter is particularly important in the genomic context, where regression analyses typically involve large sets of correlated covariates (e.g. SNPs, CNVs, transcripts). Advanced MCMC strategies were also used in the search for partition models of high dimensional associations, which arise in the multiple outcomes mapping context [5], [16]. To make our BVS strategy feasible for a large number of covariates, we exploit Graphics Processing Unit (GPU) parallelization tools and accelerated linear algebra libraries [17], which enable efficient evaluation of the marginal likelihood of millions of alternative models during the search process. An R package R2GUESS, which implements GUESS, can be downloaded from <http://www.bgx.org.uk/software/guess.html> and will soon be available on CRAN.

The SNPTEST package incorporates a Bayesian measure of association through the computation of a BF to quantify the evidence for association between a single explanatory variable and one or several continuous phenotypes. The benefits in terms of interpretability of using BFs rather than frequentist p-values in GWAS have been discussed in a number of papers [18], [19]. As SNPTEST can analyse both single and multiple traits, we will be able to compare directly the results provided by SNPTEST with those obtained by GUESS in both cases. However, SNPTEST is

Author Summary

Nowadays, the availability of cheaper and accurate assays to quantify multiple (endo)phenotypes in large population cohorts allows multi-trait studies. However, these studies are limited by the lack of flexible models integrated with efficient computational tools for genome-wide multi SNPs-traits analyses. To overcome this problem, we propose a novel Bayesian analysis strategy and a new algorithmic implementation which exploits parallel processing architecture for fully multivariate modeling of groups of correlated phenotypes at the genome-wide scale. In addition to increased power of our algorithm over alternative Bayesian and well-established non-Bayesian multi-phenotype methods, we provide an application to a real case study of several blood lipid traits, and show how our method recovered most of the major associations and is better at refining multi-trait polygenic associations than alternative methods. We reveal and replicate in independent cohorts new associations with two phenotypic groups that were not detected by competing multivariate approaches and not noticed by a large meta-GWAS. We also discuss the applicability of the proposed method to large meta-analyses involving hundreds of thousands of individuals and to diverse genomic datasets where complex dependencies in the predictor space are present.

limited to the analysis of one SNP at a time and the prior structure on the regression coefficients is less flexible than GUESS in which the data-dependent level of shrinkage conforms better to different variable selection scenarios.

Advantages over alternative GWAS non-Bayesian methods

Penalized regression methods have been proposed to improve Ordinary Least Squares, which often do poorly in both prediction and interpretation, and is not applicable in the “large p , small n ” framework. These techniques tend to shrink the regression coefficients towards zero in order to select a sparse subset of covariates and provide better prediction performance. Such methods include, among others: LASSO [20], SCAD [21], Elastic Net [22], Adaptive LASSO [23] and Fused LASSO [24].

Recently, the LASSO-type approach has been successfully applied to GWAS [25]. However, the LASSO tends to over select superfluous predictors and is not consistent for variable selection [26]. Another limitation of the original LASSO algorithm is that it cannot prioritize the most important SNPs to be selected within a group of highly correlated SNPs [22]. Improvements have been proposed such as the Smoothed Minimax Concave Penalty method [27] which accounts for the natural ordering of the SNPs and adaptively incorporates Linkage Disequilibrium (LD) information between neighboring SNPs, providing a measure of association through a resampling technique. However, such improvements are not yet implemented in LASSO-type methods for multiple phenotypes.

Building on well-established dimension reduction techniques, Sparse PLS (SPLS) [12] seeks the best linear combination of SNPs to predict a multivariate outcome of interest. The PLS approach sequentially defines components that are constructed as a linear combination of a set of predictors such that the variance explained is maximized. To ensure sparsity, the number of components to retain as well as the number of SNPs to select in each component are constrained by a penalty function on the loadings coefficients.

While both penalized regression and SPLS approaches offer solutions for multivariate GWAS, their use requires a preliminary calibration of the penalty parameters which directly affects the number of selected variables, the value of the regression coefficients and therefore the statistical performances of the models. Calibration procedures usually involve the minimization of the mean square error of prediction through V-fold cross validation. Based on the publicly available implementation of these algorithms, such procedures become computationally expensive when GWAS data are analyzed (see Material and Methods). Moreover, none of the available implementations of the aforementioned algorithms provide a measure of uncertainty of the SNP(s)-trait(s) associations. While resampling techniques could be employed [28], these would dramatically inflate the computational time. For further discussion and comprehensive comparisons of these methods, see the Power Comparison section.

Multi-phenotype analysis strategy

Beyond the methodological and computational advances of GUESS, one novel aspect of our method is the analysis strategy for groups of correlated phenotypes. This is illustrated in a study of a group of traits linked to lipid metabolism from GHS, where five lipid-related parameters Apolipoproteins A1 (APOA1) and B (APOB), HDL-cholesterol (HDL) and LDL-cholesterol (LDL) and Triglycerides (TG), are measured in 3,175 unrelated individuals [29] (see Material and Methods). The largest GWAS meta-analysis for blood lipids to date used standard single SNP analysis in a large population sample of >100,000 individuals [2]. Despite the relatively small sample size of the GHS, using our strategy, we were able to confirm the major findings reported in the GWAS meta-analysis [2] (referred to as meta-GWAS subsequently) as well as show enhanced interpretability of the results.

As illustrated in Figure 1, our strategy compares SNP-trait associations from different single and multiple phenotype combinations, starting from a meaningful phenotypic group and going down to single traits. We do not carry out a blind exploration of all possible groupings of the five traits but instead exploit the extensive biological knowledge on lipid metabolism to define two interpretable “tree like” structures. The top of the trees consist of two groups of multiple traits, TG-LDL-APOB and TG-HDL-APOA1, reflecting two main lipid metabolism pathways: the LDL (Figure 1A) and HDL pathway (Figure 1B). Considering apolipoprotein levels jointly with the lipid contents of lipoproteins may provide a more detailed insight into the lipid metabolism, the role of APOB in LDL and APOA1 in HDL, and can help elucidate the common (or specific) genetic regulation of these traits.

Our strategy is to run GUESS on the phenotypic groups at the top of each tree and on all derived subsets of traits. To compare the results between the different branches of the trees, we propose a new measure for SNP-trait(s) association, the Ratio of Bayes Factors (RBF) (see Material and Methods), to pinpoint the specific contribution of each SNP to different combinations of traits. For each SNP, by ranking the strength of association with phenotypic groups, the $\log_{10}(\text{RBF})$ allows to identify the strongest SNP-trait(s) associations and thus better characterise the biological function of the SNP on the associated trait(s).

In this study, we propose an efficient algorithm that combines the best features of genome-wide multi-SNP analysis with a fast and efficient algorithmic implementation based on Complete Unified Device Architecture (CUDA), which is extended to the analysis of multiple traits. A distinctive benefit provided by GUESS is the ability to perform a fully Bayesian analysis in an ultra-high dimensional model space and to select the best set of SNPs that predict the joint variation of several traits, which can

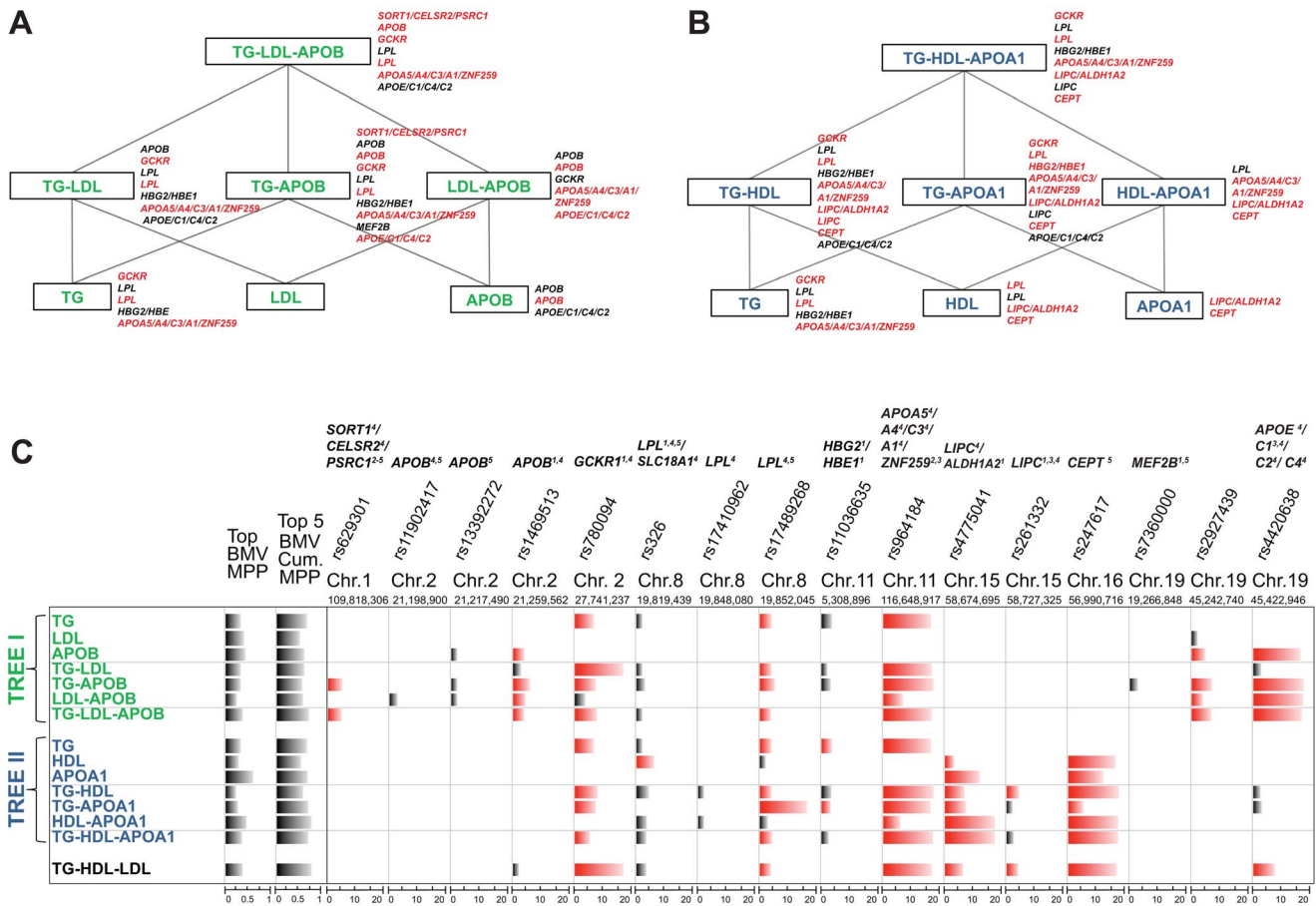


Figure 1. Schematic representation of the analysis of single and multiple phenotypes using GUESS. (A–B) Given a group of single traits (APOA1, APOB, HDL, LDL and TG), we constructed two top-down trees (green and blue colour coded) made by biologically driven combinations of phenotypes and centred on the pathways of LDL (A) and HDL (B). Each branch of the trees was regressed on the whole set of tagged SNPs (~273K SNPs) using GUESS and adjusting for sex, age and body mass index. (C) Output from GUESS is used to derive the Best Models Visited (BMV), i.e. the most supported multivariate models, and their Model Posterior Probability (MPP), i.e. the fraction of the model space explained by the BMV (MPP of the top BMV and the cumulative MPP of the top five BMV are indicated in the first two columns, respectively). Based on an empirical FDR procedure, we selected a parsimonious set of significant SNPs (indicated on the top of the table with the associated locus) that explains the variation of each branch of the two trees. Merging this information with the list of SNPs in the top BMV allowed us to highlight a robust subset of significant SNPs that repeatedly contribute to the top supported model (significant SNPs are depicted in black whereas significant SNPs that are also in the top BMV are indicated in red). For each SNPs, comparison of the marginal strength of association across different combinations of traits is possible by a new rescaled measure of marginal phenotype-SNP association, Ratio of Bayes Factors (RBF) (phenotype-SNP log₁₀(RBF) is truncated at 20 to increase readability). Based on Ensembl R66 annotation, each locus is classified as: (1) intronic, (2) 3'UTR, (3) downstream, (4) previously associated and (5) a tagSNP of a previously associated SNP. The name of the locus is also reported on the right of each branch of the two trees with the same colour code used in the table: black if the locus is associated with the phenotypes with FDR<5%, red if the locus is also in the top BMV with FDR<5%. doi:10.1371/journal.pgen.1003657.g001

provide direct insights into the polygenic regulation of multiple phenotypes.

Results

Despite the relatively modest sample size of the GHS, we were able to recover eight out of the nine top loci associated with combinations of blood lipid phenotypes that were identified by a large meta-GWAS of blood lipids in >100,000 individuals [2]: *SORT1* (rs629301), *APOB* (rs1469513), *GSKR* (rs780094), *LPL* (rs326), *APOA5* (rs964184), *LIPC* (rs261333), *CETP* (rs247617), *APOC1* (rs4420638). The only gene not detected by our approach in any combination of phenotypes was *LDLR*. This is most likely due to the lack of genotype data covering the 5'UTR of the gene where the genetic associations were previously detected (data not shown).

Enhanced interpretability of multi-phenotype associations

The multiple phenotypes approach allowed us to detect SNPs involved in combinations of traits that would have been missed by single trait analysis. For example, Figure 1C shows that rs629301, previously associated with LDL by the meta-GWAS (and Total Cholesterol (TC) as a second trait), is detected here only when considering the joint phenotype TG-APOB or TG-LDL-APOB, but surprisingly not when TG-LDL is analysed. Functional studies have shown that the causal gene responsible for lipid variations at this locus is *SORT1* which encodes sortilin, an intra-cellular receptor involved in the processing of APOB-containing particles and modulating hepatic secretion of VLDL, the lipoproteins which have the highest content of TG [30], [31]. Based on our comparative measure of association, Ratio of Bayes Factors (RBF), both TG-APOB and TG-LDL-APOB phenotypic groups

are equally associated with rs629301 by GUESS analysis (Table S1). This suggests that, besides the contribution of LDL to detect the genetic association with *SORT1*, our joint multi-trait analysis (including APOB) enhances the identification of the causal variant in this relatively small sample.

Another example relates to *LIPC* which was detected in the TG-HDL combination (and also associated with the TG-HDL-LDL group, Figure 1C and Table S1) but not with any single trait. SNP rs261333 is located within the *LIPC* gene encoding hepatic lipase which hydrolyzes TG and catabolizes TG-enriched HDL [32]. Given the tight relationship between TG and HDL in the reverse transport cholesterol pathway, considering both traits jointly enhanced the power to detect *LIPC*. In a simpler analysis, Teslovich et al. [2] looked at the marginal strongest associations with the same locus and reported the association with HDL, as a primary trait, and with TG as a secondary trait, indirectly confirming our findings.

Multi-SNP associations identified by the Best Model Visited

Figure 1C shows combinations of SNPs that have an additive effect on each phenotype or group of phenotypes. The multi-SNP association provided by GUESS Best Models Visited (BMV) enhanced the interpretation of the results and the identification of phenotypically important variants, as shown in the case of HDL and APOA1 traits (Figure 1C). APOA1 is the major apolipoprotein of HDL, and circulating levels of both traits are highly correlated (Figure S1) and are often thought to have common genetic determinants. Our multi-SNP model suggests that the main genetic locus for HDL is *CETP*, whereas both *CETP* and *LIPC* are equally involved in APOA1 determination (Table S1). This result concurs with that discussed in a recent study showing that variants in *LIPC* and *CETP* are associated with serum levels of APOA1-containing lipoprotein subfractions whereas only *CETP* is associated with HDL [33].

Another example is related to the phenotypic group TG-APOB, where the BMV enabled the identification of *GCKR* and *APOB* genes as the genetic regulators of TG-APOB in chromosome 2. Another SNP, rs13392272, which is in a non-coding region and is in high LD with rs1469513 (Figure S2), was not included in the BMV, but is only indicated as potentially marginally associated through model averaging. This highlights the ability of GUESS to differentiate variants that may not directly influence quantitative phenotypes [34]. Therefore, despite the relatively small sample size of the GHS, GUESS is able to distinguish spuriously correlated SNPs from primary associated variants.

Comparison with alternative GWAS Bayesian methods

For each branch of the two trees we compared the performance of GUESS with that of SNPTEST and for single traits with piMASS. Details about the implementation of GUESS (including the calibration of the posterior quantities) and the descriptions of SNPTEST and piMASS analysis are presented in Material and Methods.

Comparison in single trait analyses. Figures 2A–2C illustrate the genome-wide output obtained running the three algorithms for the analysis of TG trait. It is apparent how the multivariate SNPs model and the sparsity prior implemented in GUESS increase the interpretability of the results, clearly separating a small set of SNPs that are statistically associated with TG, whereas the other two plots (Figures 2B–2C) are somewhat similar and less separated. piMASS multivariate SNP model identifies the same top SNPs although the different prior specification adopted for the variance of the regression coefficients

(Table S2) leads to a less marked separation of the BF between signal and noise. In particular, a large number of SNPs had non-negligible BFs by piMASS analysis, with only small differences in BF scale between important variants and SNPs with low signal-to-noise ratio. Since piMASS does not provide the BMV, it is hard to decide if borderline associated SNPs (for instance rs17489268 and rs11036635) should be included or discarded (Figure 2C). The comparison with SNPTEST in Figure 2B shows the advantage of a multivariate SNP approach in accounting for complex LD structures. For instance, Figure 2D magnifies a region of chromosome 11 around rs964184 spanning nearly 500 Kb where SNPTEST identifies four additional SNPs connected through a complex LD pattern (rs3741298, rs6589567, rs7396835 and rs5128) that are medium/weakly correlated with rs964184. When the effect of rs964184 was removed (using standard single linear regression) none of the four additional SNPs were called significant by SNPTEST ($\log_{10}(\text{BF}) > 5$) [19]. A recent study [35] shows that haplotype associations of seven reported significant GWAS SNPs (lying from *ZNF259* to *SIK3*) with TG disappears after including rs964184 in the model, confirming the results obtained with GUESS. Figure 2E shows that the majority of SNPs detected by SNPTEST with medium/large BF are correlated (directly and/or indirectly through another SNP) with the significant SNPs found by GUESS.

Figures S3A–S3B summarise the comparison between GUESS, SNPTEST and piMASS for all the single trait analyses where, for each SNP, the genome-wide BFs of SNPTEST-GUESS and piMASS-GUESS algorithms are plotted. Overall SNPTEST is not able to separate clearly primary/secondary associations from the large bulk of SNPs (Figure S3A). There is a good agreement of the BF levels between GUESS and piMASS (Figure S3B). However GUESS outperforms the C++ version of piMASS computationally: GUESS is about 2.5 times faster than piMASS in evaluating three times more models (Table S3). Apart from the CUDA implementation of GUESS (see Material and Methods), the good performance of GUESS depends also on the prior specification of the variance of the regression coefficients (see Material and Methods and Table S2). The latter helps the search algorithm to focus on well-supported models, to reach the BVM more quickly (Table S3) and permits the fine exploration of alternative models on regions of high posterior probability (Figures S4, S5, S6).

Comparison in multi-trait analyses. Figure 3 reports the comparison between GUESS and SNPTEST for the TG-LDL-APOB group. The multivariate SNPs model and the sparsity prior implemented in GUESS enable the algorithm to identify the important genetic control points of the joint variation of TG-LDL-APOB (Figure 3A), with the top seven SNPs ranked by their BF for belonging to the BMV. In contrast, SNPTEST (Figure 3B) is not able to separate clearly the SNPs according to their joint predictive ability, and would discard well known loci. For instance, rs17489268 (*LPL* locus) and rs1469513 (*APOB*) are not included in the list of SNPs with $\log_{10}(\text{BF}) > 5$, a conventional threshold adopted for selecting significant SNPs [19] (Table S4), while GUESS includes these two SNPs in the BMV (Table S1). Moreover the separation between SNPs obtained with GUESS facilitates the application of the empirical FDR procedure (see Material and Methods and Table S5) since the null and alternative distributions are kept well apart. The overall comparison (Figure S7) shows that, as expected for any single SNP methods, SNPTEST has difficulty clearly separating the groups of associated variants from correlated SNPs. This is particularly important for the group of SNPs that are declared significant at 5% FDR by GUESS but are not in the BVM as they are hidden inside the group of predictors correlated with top associated SNPs. GWAS

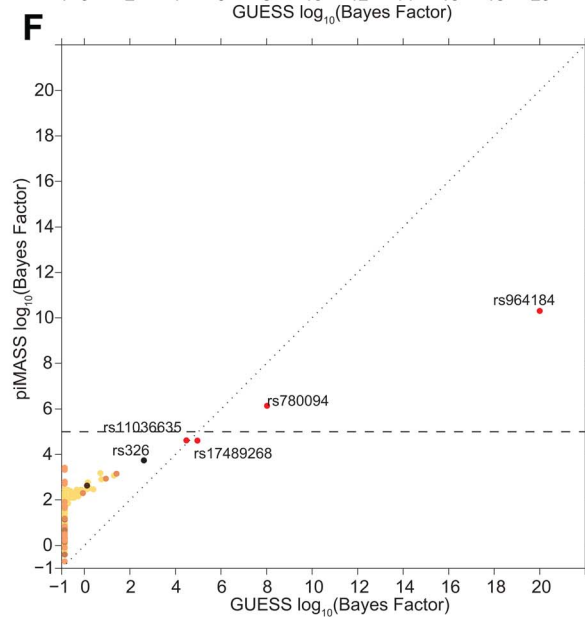
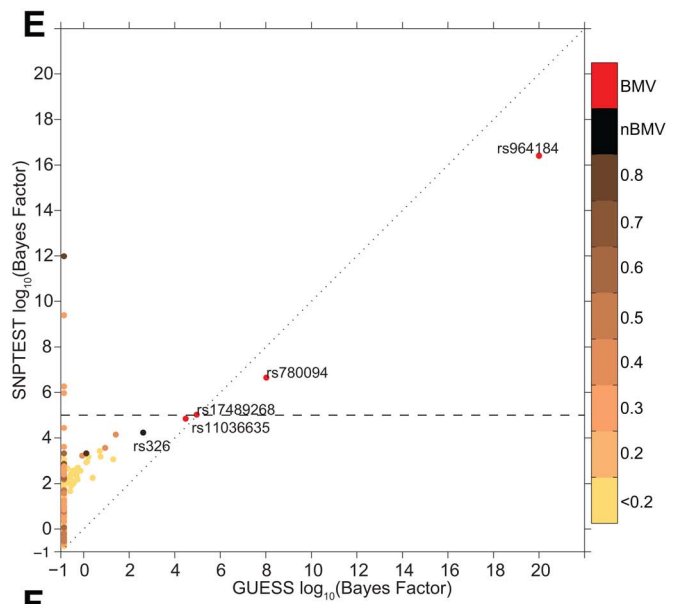
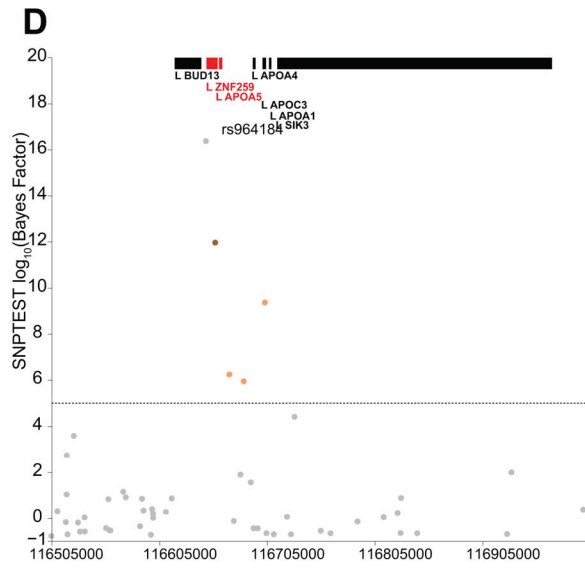
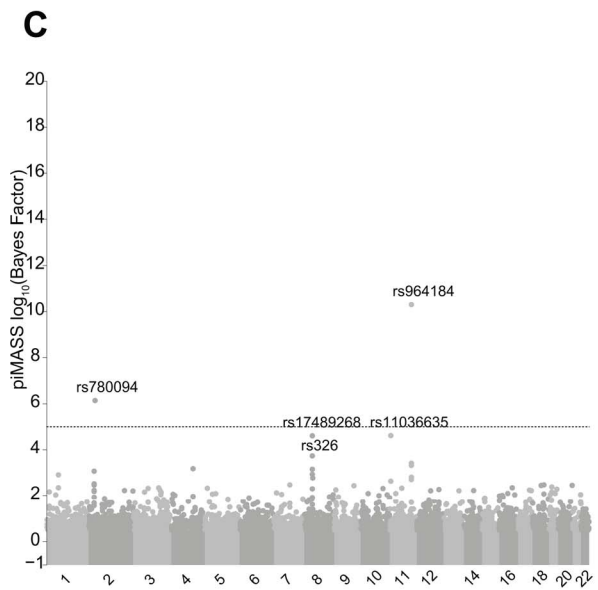
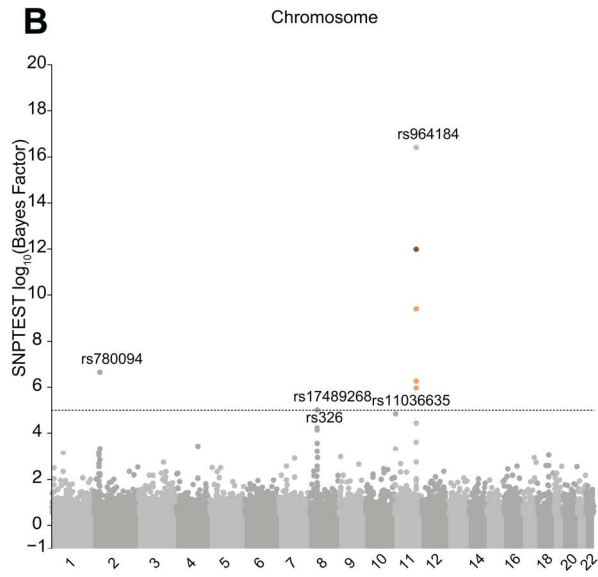
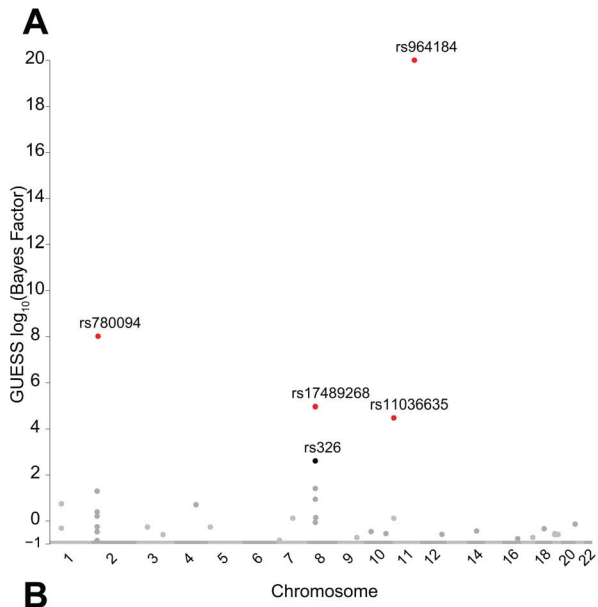


Figure 2. Comparison of the marginal phenotype-SNP associations provided by GUESS, SNPTEST and piMASS in the single trait analysis of TG. (To increase readability, the $\log_{10}(\text{BFs})$ are truncated at 20). (A) Genome-wide $\log_{10}(\text{BF})$ obtained from GUESS. Significant SNPs found associated at an FDR of 5% are depicted by black dots (with the SNP's name) whereas significant SNPs that are also in the top Best Model Visited are represented by red dots (also with the SNP's name). (B) Genome-wide $\log_{10}(\text{BF})$ obtained from SNPTEST. The horizontal dashed line indicates the level of $\log_{10}(\text{BF})$ that provides strong evidence of a phenotype-SNP association with Marginal Posterior Probability of inclusion close to 1. For comparison purposes, SNPs detected by GUESS are highlighted (their name is printed). SNPs found by SNPTEST with $\log_{10}(\text{BF}) > 5$ are coloured coded according to the level of pairwise Pearson correlation with the closest significant GUESS SNP (see colour bar for correlation scale). (C) Genome-wide $\log_{10}(\text{BF})$ obtained from piMASS. The horizontal dashed line indicates the level of $\log_{10}(\text{BF})$ that provides strong evidence for a phenotype-SNP association. (D) $\log_{10}(\text{BF})$ signals obtained from SNPTEST in a region of chromosome 11 spanning nearly 500 Kb (116,519,739–116,845,104 bp). The horizontal dashed line and colour code used to identify relevant SNPs are the same as defined in (B). Top bars indicate the position of genes in the region retrieved from Ensembl R66. (E) Scatterplot of genome-wide $\log_{10}(\text{BF})$ of TG obtained from GUESS and SNPTEST. Colour code used to identify relevant SNPs and the horizontal dashed line are as defined in (A) and (B). (F) Scatterplot of genome-wide $\log_{10}(\text{BF})$ of TG obtained from GUESS and piMASS. The colour code used to identify relevant SNPs and the horizontal dashed line are as defined in (A) and (B). doi:10.1371/journal.pgen.1003657.g002

plots for the other branches of the two trees using GUESS are presented in Figures S8, S9, S10.

Replication of multi-trait genetic associations

To demonstrate how GUESS can provide useful insights into new genetic associations with multi-phenotypes, we carried out a replication study in the Copenhagen City Heart Study (CCHS) [36], [37] and in the Data from an Epidemiological Study on the Insulin Resistance syndrome (DESIR) [38], comprising 8,261 and 4,663 individuals, respectively. We focused on two newly identified associations between *SORT1* with TG-APOB and *LIPC* with TG-HDL phenotypic groups to illustrate the added value provided by multi-trait analyses to uncover common genetic regulation underlying correlated phenotypes. To replicate both the genetic association and the order of association between the causal SNPs and the phenotypic groups we have used a two-step procedure: (1) identification of the most significant variant associated with TG-APOB and TG-HDL in each independent cohort and (2) investigation of the order of association between the variants detected in (1) and the branches of the two trees in the candidate regions.

In the first step, we selected a 2Mb region centred at each identified variant (rs629301 and rs261333) and ran GUESS in each region with an adapted specification of the *a priori* expected model size (number of true associations) and standard deviation such that the prior model size is likely to range from 0 to 3. Table S6 shows that for the selected phenotypic groups that were significantly associated with rs629301 and rs261333 in the original discovery dataset, the associations are confirmed in the two independent replication datasets. Remarkably, in CCHS and DESIR, GUESS detects the same causal variant originally identified (rs629301) for both phenotypic groups (TG-APOB and TG-LDL-APOB). The second SNP, rs261332 inside the *LIPC* gene, is not present in the CardioMetabochip [39] used for CCHS and DESIR. The variants identified by GUESS for both phenotypic groups (TG-HDL and TG-HDL-LDL) are rs8034802 (CCHS) and rs1077834 (DESIR) with r^2 and D' level equal to 0.582 and 0.979 between rs261332 and rs8034802 respectively, and 0.838 and 0.982 between rs261332 and rs1077834 respectively, in populations of European ancestry (1000 Genome project [40]). These results show that significant and novel multi-trait genome-wide associations obtained by GUESS are robust and reproducible in independent cohorts despite the relatively small size of the discovery dataset ($n = 3,175$).

In the second step, we investigated whether we would find similarities between the order of association obtained previously between the causal SNPs and the phenotypic groups (Table S1) and that obtained by applying our measure of association, RBF on the replication datasets. Specifically, for all subsets of traits in the two trees, we calculated the RBF (see Material and Methods) for

the SNPs identified in the first step as associated in each selected region (Table S6). Table S7 shows the results of this analysis for the two independent cohorts. Conditionally on rs629301, in CCHS the two phenotypic groups that receive higher RBF are TG-APOB and TG-LDL-APOB (Table S7A). The same analysis applied to the DESIR dataset (conditionally on the top BF hit SNP rs629301) provides similar results with TG-APOB ranked first (Table S7B), but with TG-LDL-APOB (ranked third) superseded by LDL-APOB. In both cases LDL is not the primary trait associated with the identified genetic variant, refining the suggested association found in [2]. In summary, the results obtained in two independent cohorts are consistent to those seen in the discovery dataset (Table S1) with the multi-trait group TG-APOB more tightly linked to the rs629301 genetic variant than any single trait. In the second region centered on rs261332, we also replicated the order of association of the phenotypic groups with rs8034802 in CCHS and rs1077834 in DESIR (Table S7A and Table S7B, respectively). In particular, in both datasets the TG-HDL-LDL and TG-HDL group receive substantially higher RBF than any other single and multiple traits group. Moreover the pattern of the RBF values is similar to that shown in the original discovery dataset (Table S1) confirming that LDL does not increase power to detect the causal variant.

Power comparison (multiple and single-trait analyses)

The real data analysis shows that SNPTEST has good power to detect the main variants but it includes several additional SNPs, possibly increasing the number of false positives. Using 273,294 SNPs from the GHS study (see Material and Methods) we carried out two simulation studies for single and multiple traits to quantify the power of SNPTEST and GUESS. In the multiple traits scenario, we also tested the performance of non-Bayesian multiple traits algorithms MANOVA [10], MLASSO [11] and SPLS [12] (see Material and Methods). We also tried a recently proposed generalised Group Fused LASSO [41], a multilocus sparse regression model which is designed to borrow information across correlated phenotypes. However the GFLASSO C++ implementation was not able to handle the whole GHS genotype dataset, requiring >33 GB RAM, while the analysis of one replicate on the smallest chromosome with cross-validation took >400 hours. For these reasons we decided to drop the comparison with GFLASSO in the simulation study. Finally, we ran GUESS with the same prior specifications used in the real data analysis (see Material and Methods), but we reduced the number of iterations to 55,000 sweeps, with 5,000 sweeps as burn-in, since the number of sweeps used in the real case study was larger than what would be required to explore adequately the posterior model space (see Material and Methods).

Multiple-trait simulation study. We simulated a group of three traits choosing four chromosomes (2, 11, 16, and 18) and, for

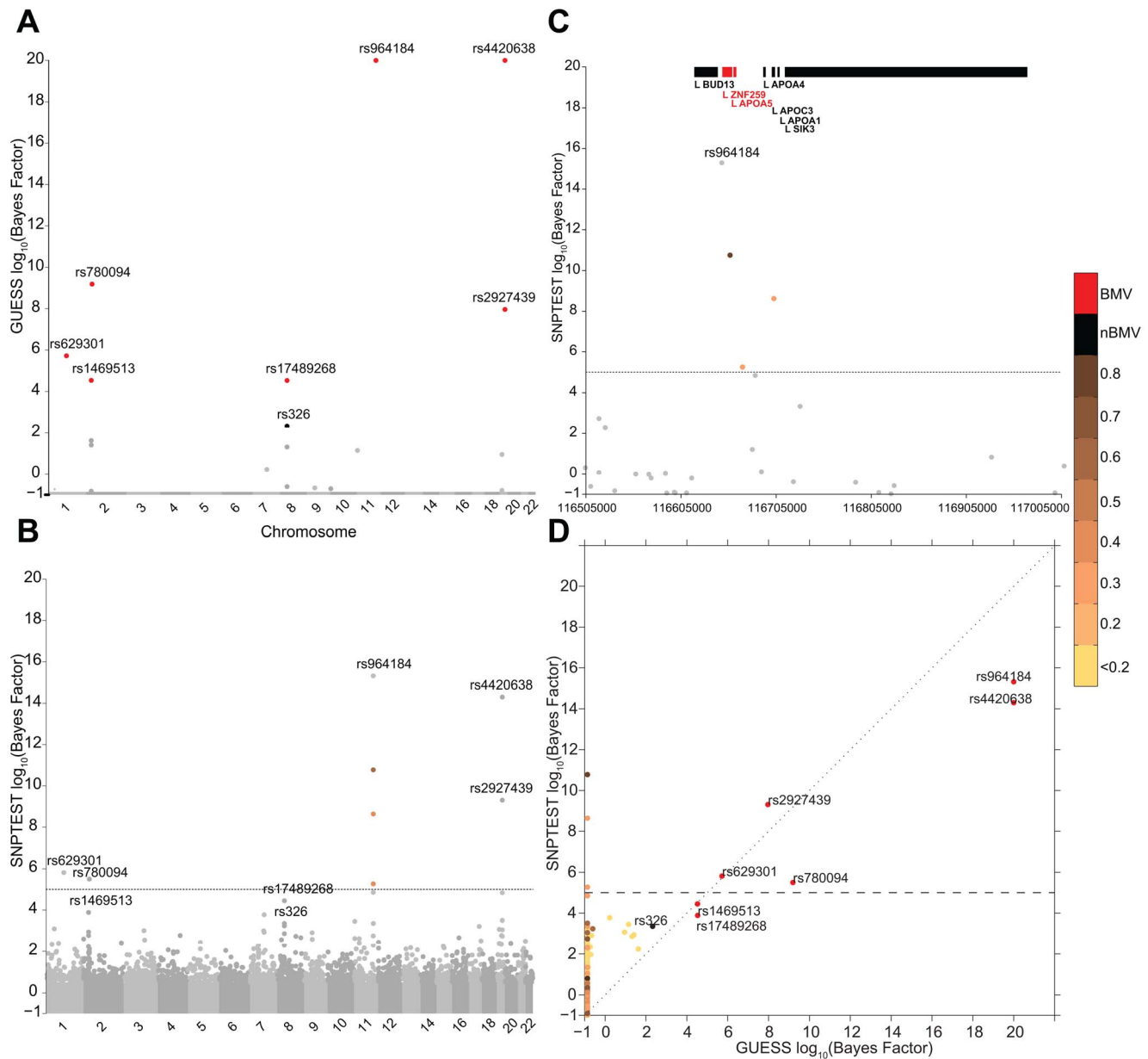


Figure 3. Comparison of the marginal phenotype-SNP associations provided by GUESS and SNPTEST in the multiple traits analysis of TG-LDL-APOB. (To increase readability, the $\log_{10}(\text{BFs})$ are truncated at 20). (A) Genome-wide $\log_{10}(\text{BF})$ obtained from GUESS. Significant SNPs found associated at 5% FDR are depicted by black dots (with the SNP's name) whereas significant SNPs that are also in the top Best Model Visited are represented by red dots (with the SNP's name). (B) Genome-wide $\log_{10}(\text{BF})$ obtained from SNPTEST. The horizontal dashed line indicates the level of $\log_{10}(\text{BF})$ that provides strong evidence of a phenotype-SNP association with Marginal Posterior Probability of inclusion close to 1. For comparison purposes, SNPs found by GUESS are highlighted (their name is printed). SNPs with $\log_{10}(\text{BF}) > 5$ are coloured coded according to the level of pairwise Pearson correlation with the closest significant GUESS SNP (see colour bar for correlation scale). (C) $\log_{10}(\text{BF})$ signal obtained from SNPTEST in a region of chromosome 11 spanning nearly 500 Kb (116,519,739–116,845,104 bp). The horizontal dashed line and colour code used to identify relevant SNPs are as defined in (B). Top bars indicate the position of genes in the region retrieved from Ensembl R66. (D) Scatterplot of genome-wide $\log_{10}(\text{BF})$ of TG-LDL-APOB obtained from GUESS and SNPTEST. The colour code used to identify relevant SNPs and the horizontal dashed line are as defined in (A) and (B).

doi:10.1371/journal.pgen.1003657.g003

each of them and in each replicate, we selected at random without replacement two SNPs. The number of SNPs selected reflects the average number of associations (7.6) found in the multiple traits real data analysis. The effects of the SNPs on the three traits were fixed ($[0.2, 0.1, 0.2, 0.1, 0.075, 0.1, 0.075, 0.1]^T$, $[0.1, 0.075, 0.1, 0.075, 0.1, 0.2, 0.1, 0.2]^T$, $[0.075, 0.1, 0.075, 0.1, 0.2, 0.1, 0.2, 0.1]^T$, respectively), but we adjusted the error variance of each trait

such that the expected proportion of variance explained is not greater than 5%. Given the effects and error variance of each trait, we simulated 20 replicates using a Normal matrix-variate distribution [42]. The residual correlation between traits was set to 0.95, 0.50 and 0.30 between the first and the second, the second and the third and first and the third trait, respectively. In a second scenario, we retained the previous setup, but we halved the

residual correlation between traits to test the multivariate methods in this more challenging case where the correlation pattern among traits is weak.

For the first simulated scenario, the Receiver Operating Characteristic (ROC) curves in Figure 4A demonstrate that, at the same Type-I error level, GUESS has higher power than SNPTTEST. When we relax the definition of false positive

associations for SNPTTEST (i.e., considering a single association in an interval centred at each top hit and spanning 25 Kb, 50 Kb and 100 Kb on both sides) the SNPTTEST ROC curves are still dominated by our GUESS multi-SNP approach (Figure S11A). When compared with non-Bayesian multiple responses methods, GUESS shows higher power than robust MANOVA over a range of FDR levels [43] and SPLS for different choices of the number of

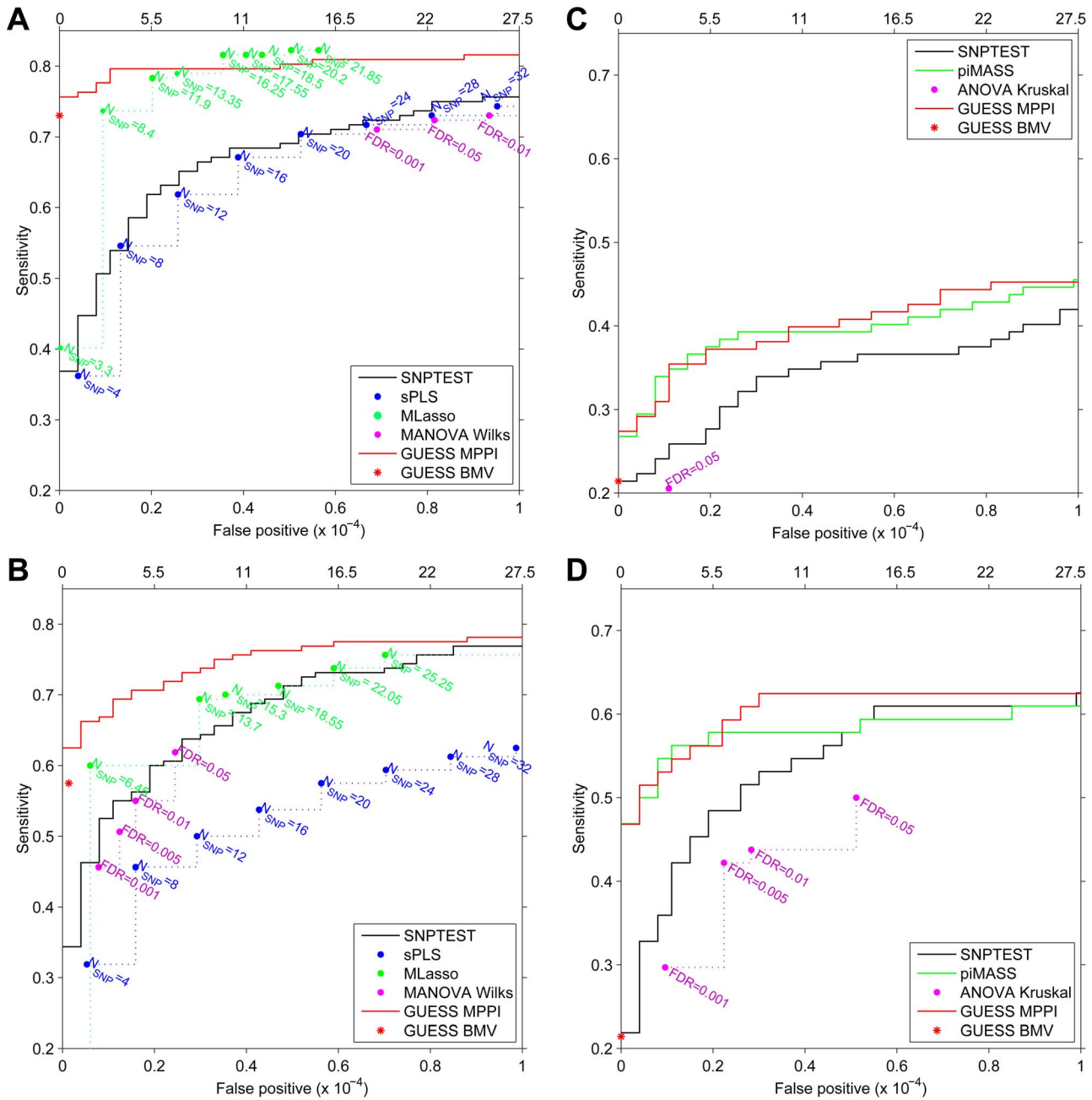


Figure 4. Receiver Operating Characteristic (ROC) curves of SNPTTEST (black), SPLS (blue), MLASSO (dark green), (M)ANOVA (purple), piMASS (green) and GUESS (red) for multiple traits and single trait simulated datasets. For GUESS, ROC curves are obtained using the top Best Model Visited (BMV) (red star) and the Marginal Posterior Probability of Inclusion (MPPI) (solid red line). For SNPTTEST, the ROC curve is calculated using $\log_{10}(\text{BF})$ while for piMASS ROC curves are obtained using MPPI. (Average) number of SNPs retained by SPLS and MLASSO under different levels of penalization are indicated (A–B). For MANOVA Wilks (A–B) and ANOVA Kruskal (C–D), the ROC curve is derived using the SNPs declared significant over a range of FDR levels. Number of false positives (x -axis) is indicated at the top of the figure while proportion of false positives is presented at the bottom. Given the large number of predictors (273,294), false positives are truncated at 10^{-4} at which level a large number already occurs (27.5).

doi:10.1371/journal.pgen.1003657.g004

SNPs retained (see Material and Methods) and when the definition of positive associations is relaxed (Figure S11C). MLASSO has slightly higher power than GUESS when the average number of SNPs detected across replicates (see Material and Methods) is larger than 16. However, in our real case study we did not notice any multiple-trait associated with more than 11 SNPs (and on average 7.6). Under this constraint, GUESS outperforms MLASSO especially when the number of false positives is low.

For the second simulated scenario, the ROC curves are depicted in Figure 4B. The power comparison between GUESS and Bayesian and non-Bayesian multiple traits methods, confirms that our algorithm also has higher power than any other method considered when the residual correlations among traits is weak. Figures S11B–S11D display the power of SNPTEST, MANOVA and SPLS when the definition of positive associations is relaxed. Also in this second scenario, GUESS BMV has higher power than any of the alternative methods investigated.

The computational time for GUESS for both multiple traits scenarios and 55000 sweeps is on average around 84 hours while MLASSO and SPLS (if cross-validation is performed) took about twice and 12 times more CPU time than GUESS, respectively.

Single-trait simulation study. Similar results are obtained when SNPTEST and GUESS are tested on a single trait. Figure 4C shows that GUESS outperforms SNPTEST when the ROC curve is calculated on the first trait of the multiple-trait first simulated scenario. GUESS also provides better results when compared with the non-parametric ANOVA test over a range of FDR levels and when the definition of positive associations is relaxed (Figure S12A). The comparison between Figure 4A and Figure S12A highlights the importance of jointly analyzing correlated multiple traits since the power to detect important variants is greatly enhanced in the multi-trait case.

The single-trait scenario allows us to also compare GUESS with piMASS. Since in Figure 4C the two methods show nearly identical power, we simulated a more complicated single-trait scenario where a secondary effect is placed close to the main effect. Specifically, we chose four chromosomes (2, 11, 16, 18) and, for each of them and in each of the 20 replicates, we selected one SNP at random. For each chromosome the second associated SNP was then selected at random from among the SNPs within 25 Kb from the first SNP. Four groups with a large and small effect that mimic primary/secondary effects ($[4, 1, 1, 6, 1.5, 3, 4, 0.5]^T$) were used to simulate the trait, adjusting the error variance such that the expected proportion of variance explained was not greater than 5%. Figure 4D shows that in this scenario GUESS and piMASS also have similar power with slightly better performance from GUESS at larger Type-I error rates. Closer inspection of the results reveals that both methods identify the majority of primary genetic associations, but GUESS was also able to detect additional SNPs with small effects. In this second single-trait scenario GUESS also outperforms SNPTEST and ANOVA over a range of FDR levels and when the definition of positive associations is relaxed (Figure S12B).

Discussion

As large scale GWAS and meta-analyses of multiple continuous phenotypes are becoming increasingly common, there is a mounting need to develop models and computationally efficient algorithms for joint analysis of multi-SNP and multi-phenotype data. Current state-of-the-art Bayesian approaches have limitations either in the analysis of one SNP at a time [9] or in modelling single phenotypes with multiple SNPs [8]. To address both these problems, we propose a powerful Bayesian statistical

computational tool for analysing genome-wide scale datasets that deals with both multiple continuous traits and predictors, with a parallelized implementation. Our algorithm enables the identification of additive effects of many predictors on multiple combinations of traits as well as secondary genetic associations. To detect multiple associated variants, stepwise-like methods have been proposed [44] but these suffer from known problems of instability when faced with correlated predictors in a high dimensional predictor space [45]. Penalised regression methods [11] and dimension reduction techniques [12] offer solutions for single and multiple-trait GWAS analysis. However since they require the calibration of the penalty parameters, they can become computationally expensive when large data are analyzed (as illustrated in the simulation study) or when resampling techniques are used to quantify the uncertainty of the SNP(s)-trait(s) associations. An alternative strategy to account for the uncertainty inherent in the model selection process is to perform model averaging [46]. This is implemented in our algorithm, GUESS, which employs the Bayesian framework for feature selection and, in particular, has the benefit of robustness and ease of interpretation of multiple SNP-trait(s) association results.

We integrated the GUESS algorithm with a new strategy for multiple traits analysis and applied this to study lipid metabolism in the Gutenberg Health Study (GHS). Despite the relatively small sample size of the GHS ($n = 3,175$) as compared with recent meta-GWAS of blood lipids [2] ($n > 100,000$ individuals), we were able to recover eight out of the nine previously identified top associations. In particular, we were able to elucidate the associations between the *SORT1* gene and the TG-APOB phenotypic group and uncover the association of *LIPC* with the TG-HDL group, which would have a low threshold of evidence if an alternative GWAS single SNP Bayesian method was used. By simply contrasting p -values for the four single traits considered and ranking them, Teslovich et al. [2] identified HDL as the leading associated trait with *LIPC* and TC as the second associated trait. Our new finding of the association of *LIPC* with multiple traits rather than with a single phenotype is supported by recent data [32]. We validated this finding in two independent datasets and, specifically, we were able to replicate the genetic association and reproduce the order of the strength of association of the genetic variant with the phenotypic groups.

Beyond the application to lipid metabolism in GHS, the strategy we propose can be applied to any set of phenotypes where unsupervised clustering methods can be used to create informative groups of traits from which a “tree-like” structure can be derived.

The increased power of GUESS shown in the real-case analysis was also demonstrated by an extensive simulation study, highlighting how alternative approaches, both Bayesian and non-Bayesian and in particular those specifically designed to deal with correlated predictors (MLASSO), are influenced by complex LD structures in the SNP data, and as a consequence have increased false positive association rates. The latter complicates, and often masks, the identification of secondary variants that are truly associated with multiple correlated traits. In contrast, the ability of GUESS to separate causal SNPs from correlated SNPs facilitates the application of empirical FDR procedures to declare robustly associated SNPs, which improves the reproducibility of results provided by GUESS.

Our implementation of BVS for high dimensional genome-wide data was made possible using the parallel computing power of the GPU interface and accelerated linear algebra libraries. In this paper we demonstrated that, exploiting the power of GPU processing, it is now feasible to run sophisticated Bayesian search algorithms in very high dimensional spaces, opening the path

towards more complex model searches that might include interaction terms. On-going work in several bioinformatics and statistical groups (<http://www.oxford-man.ox.ac.uk/gpuss/>) is fast advancing in this area and our modular algorithm will be able to benefit from these developments. One important factor in the processing speed is the number of subjects involved in the analysis. Large meta-analyses nowadays often involve hundreds of thousands of subjects and running GUESS with such a large number of individuals will be relatively slow even with new GPU implementations in the future. On the other hand, it will be feasible and relatively straightforward to use Bayesian evidence synthesis methods [47] to combine outputs from independent GUESS runs in each individual study.

In summary, we have developed a new efficient algorithm for genome-scale analysis of multiple phenotypes that maximizes genetic variants discovery and reduces complex genetic associations into understandable patterns to improve biological interpretation of results. In contrast to existing methods, the flexible prior structure used for the regression coefficients adapts to any correlation structure of the predictors, which can be of a different nature. Therefore, GUESS can be employed for large-scale analysis of multiple continuous traits with both discrete and continuous predictors and their combinations. Beyond the straightforward application to GWAS of multiple traits, GUESS is particularly suitable for the analysis of diverse genomic datasets where complex dependencies in the predictor space are present (for instance, correlation between expression levels or methylation profiles).

Materials and Methods

Samples, genotyping and traits in the primary discovery dataset

More details about the GHS study are provided in [29]. The present study included 3,175 individuals of both sexes aged 35–74 years, who were successively enrolled into the GHS, a community-based, prospective, observational single-center cohort study in the Rhein-Main region in western mid-Germany. Fasting Apolipoprotein A1 (APOA1) and B (APOB), HDL-cholesterol (HDL) and LDL-cholesterol (LDL) and Triglycerides (TG) were measured on an Architect c8000 by commercially available tests from Abbott (<http://www.abbottdiagnostics.de>). APOB is the primary apolipoprotein of LDL whereas APOA1 is the major protein component of HDL. Genotyping was performed using the *Affymetrix* Genome-Wide Human SNP Array 6.0 and the Genome-Wide Human SNP *NspI/StyI* 5.0 Assay kit. Genotypes were called using the *Affymetrix* Birdseed-V2 calling algorithm. SNPs with a Minor Allele Frequency (MAF) < 0.01 or deviating from Hardy-Weinberg equilibrium (p -value < 10^{-4}) were excluded. Only autosomal SNPs were considered for analysis.

Missing values for each of the 22 autosomal chromosomes were imputed using FastPhase [48], allowing 20 random starts of the EM algorithm ($-T20$), 100 iterations of the EM algorithm for each random start ($-C100$), no haplotype estimation ($-H-4$), without the determination of the number of clusters ($-K1$).

To reduce the number of SNPs prior to analysis, we performed tagging at $r^2 > 0.80$ level using an in-house method similar to [49]. The original dataset consisting of 650,010 SNPs was reduced to 273,294 SNPs after tagging (57.9% reduction).

Replication datasets

The Copenhagen City Heart Study [36], [37] (CCHS) is a prospective study of the Danish general population initiated in 1976–78 with follow-up examinations in 1981–83, 1991–94, and

2001–03. Individuals ($n = 8,261$) were selected based on the National Danish Civil Registration System to reflect the adult Danish population aged 20–100 years. Data were obtained from a questionnaire, a physical examination, and from blood samples including DNA extraction at the 1991–94 examination. A lipid profile was measured using standard hospital assays and genotyping was performed using customised version of the Illumina CardioMetaboChip [39]. For the replication, we selected a region centered at rs629301 (*SORT1*) and rs261333 (*LIPC*) comprising 543 and 204 SNPs, respectively.

We also analyzed 4,663 subjects of European descent from the Data from an Epidemiological Study on the Insulin Resistance syndrome (DESIR) cohort. More details about this study are available in [38]. The subjects were genotyped using the Illumina CardioMetaboChip genotyping array. None of those individuals were prescribed lipid lowering treatments. Serum HDL-cholesterol was assayed by the phosphotungstic precipitation method while total cholesterol and triglycerides levels were assayed by the enzymatic Trinder method. These measurements were obtained using a Technicon DAX24 from Bayer Diagnostics, Puteaux, France or using a Delta a 60i from Konelab, Evry, France. Apolipoprotein B levels were measured by nephelometry using a BNA or BN 100 nephelometer from Behring, Reuil Malmaison, France. The regions selected for replication comprise 1,003 and 442 SNPs spanning 1.94 and 1.97 Mb, respectively.

GUESS implementation for large number of predictors

The GUESS implementation extends the original ESS++ code [50], permitting an effective posterior exploration of model spaces of the size typically encountered in GWAS problems. Similarly to ESS++, GUESS simulates multiple Markov chains in parallel, with a different temperature attached to each chain. The different temperatures have the effect of flattening the log-Posterior (log-marginal likelihood \times log-prior on the model space). The state of the chains is tentatively swapped at every iteration by a within- and between-chains probabilistic mechanism. This ensures that the posterior distribution is not trapped in any local mode and that the algorithm mixes efficiently since every chain constantly tries to transmit information about its state to the others. For interested readers, description of the probabilistic swapping mechanisms, i.e. local (Fast Scan Metropolis Hastings sampler) and global moves (Crossover operators, Exchange operator) implemented in GUESS, their efficiency to explore the posterior model space as well as the automatic tuning of the temperature ladder are discussed in details in [7].

As indicated by its name, the novel implementation involves the use of Graphical Processor Unit (GPU) technologies, specifically using NVIDIA's Complete Unified Device Architecture (CUDA), <http://developer.nvidia.com/category/zone/cuda-zone>. CUDA is a parallel processing architecture that utilizes the processing power of the many processors present on a GPU, allowing significant performance increases for many mathematical operations and algorithms. By rewriting code in CUDA C/C++ parts of the algorithm can be redirected to the GPU rather than the CPU, often greatly speeding up a typical run [17]. As detailed by [7], at each MCMC update of the ESS algorithm, it is necessary to evaluate the log-Posterior, which requires the expensive computation of the marginal likelihood. To increase stability, the marginal likelihood is calculated using the technique of QR matrix decomposition, as described in [3]. For variable selection problems where the number of possible predictors in the model can be large, performing QR decomposition using regular CPU operations becomes prohibitively computationally expensive, resulting in infeasible run times. GUESS replaces core linear

algebra operations, including the QR decomposition, with versions that exploit the GPU. In the implementation used to produce the results described in this paper, we use version R11 of the proprietary CULA library (<http://www.culatools.com/>), which is freely available to academic users, directly replacing the GNU Scientific Library (<http://www.gnu.org/software/gsl>) versions of the relevant linear algebra routines present in the ESS++ code, with CUDA C/C++ equivalents from this library.

Beyond the primary extension to ESS++, GUESS also implements a slight difference in the Metropolis-Hastings move type of the underlying algorithm (see [7]). In particular, for the heated chains, the original move allowed the probability of proposing whether a particular variable was included or not to depend upon the temperature of the chain. We found that this encouraged too many proposals to models with a large number of variables (in the heated chains) which were very frequently rejected. By altering the algorithm so that the proposal probability no longer depended on the temperature of the chain (and changing the acceptance probability accordingly) the efficiency of the algorithm was improved.

Whilst the change to GPU based linear algebra routines marks a significant performance improvement, even with these changes in place, attempting to evaluate the marginal likelihood for a model with many variables can be prohibitively slow. Because we put a strong penalty on such models through the prior on the number of predictors in the model, they are typically very rarely visited by the unheated chain in the transient phase of the algorithm, when the posterior is being explored. However, in the burn-in phase of the chain, or for the heated chains, such models might be visited more frequently.

To prevent inefficiency in the burn-in phase and allowing the successful exploration of the posterior density, we truncate the prior on the number of variables in the model to exclude any models with too many variables. This truncation leads to a re-normalization of the posterior distribution, but as the normalization constant is not required in the acceptance ratio of the affected MCMC moves (local moves), in practice, the algorithm rejects any proposed moves to any model with more than the permitted number of variables.

The truncation (T) is set by the user through an additional parameter (F), through the relation $T = E + F \times S$, where E is the expected value and S is the standard deviation of the prior model size p_γ . Given the very large number of predictors (SNPs) in a GWAS and the Central Limit Theorem approximation of a binomial distribution already for moderate values of F , for instance $F > 3$, $\Pr(p_\gamma > T) = 1 - \Phi(F) \approx 0$, so that the truncation has a negligible effect. The space of candidate models is reduced from 2^p to $\sum_{j=1}^T \binom{p}{j}$ which is still considerably large.

Finally, in GUESS we use the same hierarchical conjugate prior structure for the regression coefficients presented in [7], where the g -prior on the genetic effects, that replicates the covariance structure of the likelihood, is coupled with an inverse gamma hyper-prior on the selection coefficient g , giving rise to the so-called Zellner-Siow prior and a recommended heavy tailed distribution for the regression coefficients.

The original GPU-enabled version of GUESS/ESS++ is freely available at <http://www.bgx.org.uk/software/guess.html> with an installation guide and an extensive description of the features of the algorithm. Moreover, GUESS has been wrapped into an R package called R2GUESS which provides an easy way to install and run the CULA/C++ version of the GUESS code, including an integrated post-processing of the output and automatic FDR

calculation. It can be downloaded from <http://www.bgx.org.uk/software/guess.html> and will soon be available on CRAN.

GUESS analysis

After performing normal quantile transformation for each single trait, we run GUESS for each branch of the trees shown in Figure 1 adjusting for sex, age and body mass index which were considered important for all models. We imposed sparsity with $E = 20$, $S = 12$, the *a priori* expected model size (expected number of true associations) and standard deviation of the model size, and $F = 7$, meaning the prior model size is likely to range from 0 to 56 with a maximum model size of $T = 104$. In this set-up, given the level of sparsity and the number of predictors ($p = 273,294$), the average prior probability π that a SNP is truly associated with the phenotype is 7.32×10^{-5} which is well inside the range of the prior probability suggested by [19] for Bayesian GWAS. GUESS was run for 110,000 sweeps, with 10,000 sweeps as burn-in, with three chains run in parallel (number of chains chosen after a pilot study) and a hyper-prior on the selection coefficient g . The analysis was performed on a HPC cluster computer with a 2.8 GHz Dual-Core Xeon processor and an NVidia Tesla C1060 GPU with 8 Gb of RAM. Average computational times for the single- and multi-trait analysis were 25₅² and 229 hours, respectively (Table S3 for details). Visual inspection of the trace of the log-Posterior (log-marginal likelihood \times log-prior on the model space), model size and selection coefficient g show the chains converged to their apparent stationary distributions (Figure S4 for TG-HDL-LDL group). As illustrated in Figure S5 for the TG-HDL-LDL group, GUESS is able to move very quickly towards competing models well supported by the data, highlighting the fact that the number of sweeps used is larger than would be required for a faithful exploration of the model space. Formal diagnostic tests for convergence were performed similarly to [42]. Table S3 shows for each group of phenotypes the number of models visited and the number of models explored before visiting the top BMV (after the burn-in phase), the number of unique models visited (after burn-in phase), the models average size and the overall computational time (in hours). While most of the time the BMV is visited immediately after the end of the burn-in phase, for two phenotypic groups, TG-LDL and in particular LDL-APOB, the number of models visited before reaching the BMV is quite large, suggesting that for multiple and diverse groups of traits running the algorithm for a large number of iterations is recommended in order to explore the huge model space of predictors.

To evaluate the impact of the prior setup on the regression coefficients and the choice of the hyper-coefficients of the sparsity prior, we performed a sensitivity analysis. Firstly, we implemented a new version of our algorithm based on a conjugate hierarchical independent prior for the genetic effects with a diffuse exponential hyper-prior for the variance of the regression coefficients [7]. Table S8 shows that results are very consistent with those obtained with the Zellner-Siow prior (Best Models Visited, Top BMV Posterior Probability and the Top 5 BMV Posterior Probability), suggesting that when the number of observations is large, as typically the case in GWAS, the prior structure is dominated by the likelihood [18]. Secondly, we tested the effect of the hyper-coefficients of the sparsity prior on the multiple-trait simulation study. Specifically, we halved and doubled the *a priori* expected model size, $E = 10$ and $E = 40$, respectively, while keeping the same value of the standard deviation of the model size, $S = 12$. With these two new input parameterizations, the prior model size is likely to range from 0 to 46 with a maximum model size of $T = 94$ and 0 to 76 with a maximum model size of $T = 134$, respectively. Figure S13 shows the ROC curves for the first five

replicates of the two simulated multi-trait examples under the different sparsity prior settings. Although the average prior probability π that a SNP is truly associated with the phenotype now ranges between 3.66×10^{-5} and 1.47×10^{-4} , its value is still relatively low with a negligible impact on results.

GUESS output and empirical FDR

GUESS provides two types of output. The first is the Best Models Visited (BMV), i.e. the most supported multivariate models ranked according to their Model Posterior Probability (MPP). For each multivariate model visited during the MCMC, the log-Posterior (log-marginal likelihood \times log-prior on the model space) is available and, for each unique model visited, the MPP is equal to the renormalized log-Posterior (with respect to all unique models visited). See [6] for details. The second type of output is related to the Marginal Posterior Probability of Inclusion (MPPI). As detailed in [50], MPPI provides a model-averaged measure of importance of each predictor with respect to the models visited and can be interpreted as the posterior strength of association between a single SNP and a group of phenotypes.

Several alternatives have been proposed in the literature to select significant MPPI either based on prediction considerations [51] or FDR principles [52]. Here, we proposed a strategy similar to the ‘‘Bayes/non-Bayes compromise’’ described in [53]. However, instead of deriving empirical p -values as the proportion of permuted datasets for which the MPPI exceeds the MPPI for the observed data, the permutation strategy allows us to define the MPPI threshold at a specific empirical FDR level. Specifically, for each group of phenotypes, we compute the MPPI for the observed data and, based on the same prior specification and the same parameters for the GUESS algorithm, the MPPI for artificial datasets created by permuting three times the rows (subjects) of the observed traits. Overall for each group of traits 819,882 (273,294 \times 3) observations from the null distribution were obtained using this procedure. The MPPI threshold is then defined as the MPPI level for which the ratio between the number of declared associations in the shuffled datasets and the observed dataset is not greater than a specified FDR level. Since the sample size needs to be large to evaluate the tail of the MPPI distribution in the artificial datasets, we combined the MPPI of the null distributions for all the artificial groups of phenotypes with the same dimension (triplets, pairs and singleton). Table S5 shows for each branch of the two trees the sample size (and number of artificial datasets) used to approximate the MPPI null distribution, the MPPI threshold and the number of MPPI declared significant at 5% empirical FDR.

Ratio of Bayes Factors

Bayes Factor (BF) for the j th SNP in the g th group of phenotypes is defined as

$$BF(\gamma_{jg} = 1; \gamma_{jg} = 0) = \frac{p(\gamma_{jg} = 1 | \mathbf{Y}_g)}{1 - p(\gamma_{jg} = 1 | \mathbf{Y}_g)} \bigg/ \frac{E(p_{\gamma_g})/p}{1 - [E(p_{\gamma_g})/p]} \quad (1)$$

where the numerator is the Posterior Odds and the denominator is the Prior Odds. For each SNP in the g th group, it compares two different models ($\gamma_{jg} = 1$ vs $\gamma_{jg} = 0$) regardless of the prior probability [54]. Let $p^{FDR}(\gamma_{jg} = 1 | \mathbf{Y}_g)$ be the MPPI threshold obtained from each group of phenotypes at a specified FDR level obtained through permutation. The corresponding BF threshold is

$$BF^{FDR}(\gamma_{jg} = 1; \gamma_{jg} = 0) = \frac{p^{FDR}(\gamma_{jg} = 1 | \mathbf{Y}_g)}{1 - p^{FDR}(\gamma_{jg} = 1 | \mathbf{Y}_g)} \bigg/ \frac{E(p_{\gamma_g})/p}{1 - [E(p_{\gamma_g})/p]}$$

which provides the threshold on the BF scale, at some FDR level, to call the j th SNP associated with the g th group of phenotypes regardless of the prior probability. It is expected that this threshold varies depending on the number and correlation of the group of phenotypes. The quantity

$$\frac{BF(\gamma_{jg} = 1; \gamma_{jg} = 0)}{BF^{FDR}(\gamma_{jg} = 1; \gamma_{jg} = 0)} = \frac{p(\gamma_{jg} = 1 | \mathbf{Y}_g)}{1 - p(\gamma_{jg} = 1 | \mathbf{Y}_g)} \bigg/ \frac{p^{FDR}(\gamma_{jg} = 1 | \mathbf{Y}_g)}{1 - p^{FDR}(\gamma_{jg} = 1 | \mathbf{Y}_g)} \quad (2)$$

rescales the BF with respect to its FDR ‘‘baseline’’ level obtained in each group. The Ratio of Bayes Factors (2) (with $RBF_{jg} \geq 1$ since a SNP is declared associated if $p(\gamma_{jg} = 1 | \mathbf{Y}_g) \geq p^{FDR}(\gamma_{jg} = 1 | \mathbf{Y}_g)$) is similar to (1), but there is an important difference that distinguish them: the former is the ‘‘relative measure of risk’’ of the j th SNP to be associated with the g th group of phenotypes with respect to the prior beliefs, while the latter is the ‘‘relative measure of risk’’ of the j th SNP to be associated with the g th group of phenotypes with respect to the MPPI threshold obtained from each group g at a specified FDR level. The denominator in (2) acts as a standardisation factor. For a given SNP j , RBF_{jg} can be compared across groups of traits and provides a formal way to rank them with respect to the strength of association with the SNP.

Let RBF_{jg} and RBF_{jh} be the RBF defined in (2) for two groups of traits. If $RBF_{jg} > RBF_{jh}$, then

$$\frac{BF_{jg}}{BF_{jh}} > \frac{BF_{jg}^{FDR}}{BF_{jh}^{FDR}}.$$

Therefore $RBF_{jg} > RBF_{jh}$ if the ratio of the Bayes Factors of the two groups of traits is larger than the ratio of the Bayes Factors at the FDR baseline level (that can be $<$ or > 1). Finally, given a groups of traits, if $BF_{ig} > BF_{jg}$, $i \neq j$, then $RBF_{ig} > RBF_{jg}$, showing that the new measure does not alter the rank of the phenotype(s)-SNP association within each group.

SNPTEST analysis

SNPTEST V2.2.0 (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html) automatically performs normal quantile transformation to each trait and adjusts for sex, age and body mass index (`-cov_all`). We chose the Bayesian analysis (`-bayesian 1`) with suggested default hyper-parameters for the single trait analysis (normal prior on the effect centred in 0 (`-prior_qt_mean_b 0`) and variance $0.02\sigma^2$ (`-prior_qt_v_b 0.02`)) and InverseGamma prior on the error variance σ^2 with finite mean 1 (`-prior_qt_a 3` and `-prior_qt_b 2`). For the multiple traits analysis we selected the suggested default values (normal matrix prior on the effects centred in $\mathbf{0}$ (`-prior_qt_mean_b 0`) with covariance matrix $0.02 \sum$ (`-prior_qt_v_b 0.02`)) and InverseWishart prior on the error variance \sum (`-prior_mqt_c 6` and `-prior_mqt_Q 4`). The prior probability of association cannot be modified and it is set at $\pi = 10^{-4}$. SNPTEST provides the value of the BF automatically. Table S2 compares the prior setup and hyper-priors coefficients used in GUESS and SNPTEST.

piMASS analysis

Each trait was normal transformed using the R function `qnorm`. The effect of sex, age and body mass index was removed

by performing standard multiple linear regression software and then running piMASS v0.9 (www.bcm.edu/cnrc/mcmcmc/pimass) on the residuals from this regression. In order to match the prior on the model size used in GUESS, we set the minimum and the maximum of π (the prior probability that a SNP is truly associated with the phenotype) equal to 1 and 56 out of the total number of SNPs (`-pmin 1 -pmax 56`), restricting the minimum and maximum number of SNPs in a model to be 1 and 104 (`-smin 1 -smax 104`) without any constraint on the hyper-parameter h and no cut-off on MAF (`-exclude-maf 0`). We ran piMASS with 10^6 warm-up steps followed by 10^7 sampling steps (`-w 1000000 -s 10000000`), recording a sample every 10 steps (`-num 10`). Although we did not match the number of visited models by GUESS with those of piMASS (for piMASS the number of sampling steps coincides with the number of models visited), we are confident that the very large number of sampling steps allows piMASS to explore faithfully the model space. Table S3 shows for each single trait the computational time required by piMASS to complete the task while Figure S6 presents the trace plot of the model $\log_{10}(\text{BF})$ for TG. Since piMASS provides the MPPI through Rao-Blackwellization [8], but not the BF for each SNP, we calculate it as in (2) with $E(p_{\gamma}) = 13.663$ which corresponds to $E(\pi) = 5 \times 10^{-5}$. Finally Table S2 compares the prior set-up and hyper-priors coefficients used in GUESS and piMASS.

Multivariate ANOVA analysis

We implemented the frequentist analysis of multiple traits using the function `wilks.test` from the `rrcov` R package (<http://cran.r-project.org/web/packages/rrcov/>) to compare the responses' means for each SNP in the simulated experiments. Setting `method=rank` the classical Wilks' Lambda statistic for testing the equality of the group means for all the responses is modified into a robust version [10]. For the single trait analysis we used the non-parametric ANOVA function `kruskal.test` implemented in the R package `stats`. In both cases Storey's FDR method [43] was used to control for multiple testing and to call associated SNPs. Finally, in the power calculation, the definition of false positives was relaxed by considering a single association in the interval centred at each top associated SNP with the multiple traits and spanning 100 Kb on both sides (Figures S11C–S11D and Figure S12).

Sparse SPLS analysis

We used the `spls` function from the `mixOmics` R package (<http://cran.r-project.org/web/packages/mixOmics/index.html>) [55], [56] to predict the multivariate outcome by a linear combination of SNPs. In accordance with the structure of the multiple traits simulated datasets, we only retained one axis (`ncomp=1`) and investigated nine different values of the number of SNPs in this component (`keepX`) ranging from 4 to 36. SNPs contributing to the component are defined as those with non-zero loadings coefficient. In this special case where only one component is retained for the regression model, SPLS corresponds to canonical regression [12].

Building on the known structure of the multiple responses simulated dataset we were able to fix the number of components as well as the number of the SNPs contributing to each component. The analysis of each replicate took approximately 40 minutes. Using the model on real data, these two features have to be assessed by means of a V-fold cross-validation procedure. Using standard 10-fold cross-validation replicated 50 times, each combination of `ncomp` and `keepX` would take over 33 hours. In summary, even when browsing a limited number of combinations of values for `ncomp` and `keepX`, (e.g. `ncomp` ranging from 1 to 3,

and `keepX` ranging from 1 to 100 with an increment of 10) the overall computational time required by SPLS is around 12 times greater than that of GUESS.

Multivariate LASSO analysis

We fitted a LASSO-type penalized multivariate linear regression model using `glmnet` (<http://cran.r-project.org/web/packages/glmnet/index.html>) R package [11]. The LASSO penalty used in this model generalizes the group LASSO penalty to account for potential correlation within the multivariate response. To accommodate for continuous multiple responses, the response type was set to `family=mgaussian`, and the LASSO penalty was enabled by setting `alpha=1`. The penalty λ was calibrated based on the first replicate of each multiple traits simulated dataset such that the number of retained SNPs by MLASSO is consistent with the sequence of values of `keepX` defined in the SPLS analysis. The resulting set of nine values for λ was used in all replicates of the two simulated multiple traits scenarios.

Similarly to SPLS, application of this group LASSO procedure on real data will require the calibration of λ . Running a 10-fold cross-validation procedure over a grid containing 100 values of λ replicated 50 times, would yield an average computing time exceeding 110 hours.

Supporting Information

Figure S1 Heat-map of the correlation matrix of the five traits used in the tree analysis. Off-diagonal correlation between each pair of traits is indicated inside the heat-map. (TIF)

Figure S2 Heat-map of the squared correlation matrix of the 16 SNPs which were marginally associated with any group of traits using an empirical FDR cut-off of 5%. Squared correlation between `rs11902417` and `rs13392272` is 0.2711, `rs11902417` and `rs1469513` is 0.1949 and `rs13392272` and `rs1469513` is 0.5582. Squared correlation between `rs326` and `rs17410962` is 0.3305, `rs326` and `rs17489268` is 0.7901 and `rs17410962` and `rs17489268` is 0.2165. (TIF)

Figure S3 Comparison of the marginal phenotype-SNP association provided by GUESS, SNPTTEST and piMASS for all single traits of two trees. (To increase readability, the $\log_{10}(\text{BFs})$ are truncated at 20). (A) Scatterplot of $\log_{10}(\text{BF})$ GUESS vs SNPTTEST obtained superimposing the scatterplot of each single trait. A horizontal dashed line indicates level of $\log_{10}(\text{BF})$ that provides strong evidence of a phenotype-SNP association ($\log_{10}(\text{BF}) > 5$). Red and black dots highlight significant SNPs found by GUESS while non-significant SNPs are colour coded according to the level of pairwise Pearson correlation with the closest significant GUESS SNP (see the colour bar for the correlation scale). (B) Scatterplot of $\log_{10}(\text{BF})$ GUESS vs piMASS obtained superimposing the scatterplot of each single trait. Colour code used to identify relevant SNPs and horizontal dashed line are the same as defined in (A). (TIF)

Figure S4 GUESS diagnostic plots in the TG-HDL-LDL group analysis. (A) Trace plot of the log-Posterior (log-marginal likelihood \times log-prior on the model space) of the three chains run in parallel. (B) Trace plot of the size of the models explored by the three chains run in parallel. (C) Trace plot of the selection coefficient g (blue) and shrinkage factor $g/(1+g)$. In all plots, black vertical dotted line indicates the end of the burn-in phase. (TIF)

Figure S5 Trace plot of the size of the models explored by the non-heated chain of GUESS in the TG-HDL-LDL group analysis. Letters A-H indicate when GUESS first identifies the top Best Model Visited (A), the second Best Model Visited (B) and etc. with models ranked by the Model Posterior Probability. A black vertical dotted line indicates the end of the burn-in phase.

(TIF)

Figure S6 Trace plot of piMASS “Model $\log_{10}(\text{BF})$ ” in the single trait TG analysis. Values of $\log_{10}(\text{BF})$ are recorded every ten iterations.

(TIF)

Figure S7 Comparison of the marginal phenotype-SNP association provided by GUESS and SNPTEST for all multiple traits of two trees. (To increase readability, the $\log_{10}(\text{BFs})$ are truncated at 20). Scatterplot of $\log_{10}(\text{BF})$ GUESS vs SNPTEST obtained superimposing the scatterplot of each multiple trait group. A horizontal dashed line indicates the level of $\log_{10}(\text{BF})$ that provides strong evidence of a phenotype-SNP association ($\log_{10}(\text{BF}) > 5$). Red and black dots highlight significant SNPs found by GUESS while non-significant SNPs are colour coded according to the level of pairwise Pearson correlation with the closest significant GUESS SNP (see the colour bar for the correlation scale).

(TIF)

Figure S8 Genome-wide $\log_{10}(\text{BF})$ obtained from GUESS for single trait analysis. (A) LDL, (B) APOB (first tree), (C) HDL and (D) APOA1 (second tree). Significant SNPs found associated at a 5% FDR are depicted by black dots (with the SNP’s name) whereas significant SNPs that are also in the top Best Model Visited are represented by red dots (with the SNP’s name) (the $\log_{10}(\text{BF})$ is truncated at 20).

(TIF)

Figure S9 Genome-wide $\log_{10}(\text{BF})$ obtained from GUESS in the first tree centred in the LDL pathway. (A) TG-LDL, (B) TG-APOB and (C) LDL-APOB. Significant SNPs found associated at a 5% FDR are depicted by black dots (with the SNP’s name) whereas significant SNPs that are also in the top Best Model Visited are represented by red dots (with the SNP’s name) (the $\log_{10}(\text{BF})$ is truncated at 20).

(TIF)

Figure S10 Genome-wide $\log_{10}(\text{BF})$ obtained from GUESS in the second tree centred in the HDL pathway. (A) TG-HDL, (B) TG-APOA1, (C) HDL-APOA1 and (D) TG-HDL-APOA1. Significant SNPs found associated at 5% FDR are depicted by black dots (with the SNP’s name) whereas significant SNPs that are also in the top Best Model Visited are represented by red dots (with the SNP’s name) (the $\log_{10}(\text{BF})$ is truncated at 20).

(TIF)

Figure S11 Receiver Operating Characteristic (ROC) curves comparison. ROC curves of SNPTEST (black), SPLS (blue), MANOVA (purple), and GUESS (red) for the first (A–C) and second (B–D) multiple-trait simulated datasets when the definition of positive associations is relaxed, i.e. considering a single association in an interval centred at each top hit and spanning 25 kb, 50 kb and 100 kb on both sides. For GUESS, separate ROC curves are obtained using the top Best Model Visited (red star) and the Marginal Posterior Probability of Inclusion (solid red line). For SNPTEST, the ROC curve is calculated using the $\log_{10}(\text{BF})$. The number of SNPs retained by SPLS under different levels of penalization is indicated. For MANOVA Wilks, the ROC curve is derived using SNPs declared significant over a range of FDR levels. The number of false positives (x-axis) is indicated at

the top of the figure while the proportion of false positives is presented at the bottom. Given the large number of predictors (273,294), false positives are truncated at 10^{-4} at which level a large number already occurs (27.5).

(TIF)

Figure S12 Receiver Operating Characteristic (ROC) curves comparison. ROC curves of SNPTEST (black), ANOVA (purple) and GUESS (red) for the first (A) and second (B) single-trait simulated datasets when the definition of positive associations is relaxed, i.e. considering a single association in an interval centred at each top hit and spanning 25 kb, 50 kb and 100 kb on both sides. For GUESS, separate ROC curves were obtained using the top Best Model Visited (red star) and the Marginal Posterior Probability of Inclusion (MPP) (solid red line). For SNPTEST, the ROC curve is calculated using the $\log_{10}(\text{BF})$. For ANOVA Kruskal, the ROC curve is derived using SNPs declared significant over a range of FDR levels. The number of false positives (x-axis) is indicated at the top of the figure while the proportion of false positives is presented at the bottom. Given the large number of predictors (273,294), false positives are truncated at 10^{-4} at which level a large number already occurs (27.5).

(TIF)

Figure S13 Receiver Operating Characteristic (ROC) curves of GUESS under different parameterization. ROC curve of the *a priori* expected model size, i.e. $E = 10$ (blue), $E = 20$ (red) and $E = 40$ (green) for five replicates of the first (A) and second (B) multi-trait simulated dataset are depicted. Separate ROC curves were obtained using the top Best Model Visited (star) and the Marginal Posterior Probability of Inclusion (solid line). Given the large number of predictors (273,294), false positives are truncated at 10^{-4} at which level a large number already occurs (27.5).

(TIF)

Table S1 Post-processed output obtained from GUESS for all the elements of the two trees (green and blue colour coded) and TG-HDL-LDL. Horizontal lines separating groups of traits with the same cardinality (singleton, pairs and triplets). Model Posterior Probability (MPP) of the top Best Model Visited (BMV) and the cumulative MPP of the five top BMV are indicated in the first two columns of the table, respectively. The unique set of significant SNPs ($\text{FDR} < 0.05$) which predict a group of phenotypes is indicated on the top of the table as well as the associated locus. Based on Ensembl R66 annotation, each locus is classified as: (1) intronic, (2) 3’UTR, (3) downstream, (4) previously associated and (5) a tagSNP of a previously associated SNP. In the centre of the table $\log_{10}(\text{RBF})$, i.e. rescaled marginal phenotype-SNP association, are included with significant SNPs depicted in black and significant SNPs that are also in the top BMV indicated in red (the $\log_{10}(\text{RBF})$ is truncated at 20). The Ratio of Bayes Factors (RBF) is a rescaled measure of SNP-trait(s) association and it is defined as the ratio between the BF to test the SNP-trait(s) association hypothesis and the “baseline” BF level obtained through permutations.

(PDF)

Table S2 Comparison of prior setup between GUESS, SNPTEST and piMASS. In SNPTEST the hyper-priors on π and on the variance of the regression coefficients are not specified. piMASS differs from GUESS by a different specification of the priors on the regression coefficients and on their variance.

(PDF)

Table S3 Comparison of the MCMC efficiency between GUESS and piMASS. GUESS was run for 100,000 sweeps with 10,000 as burn-in and with 3 chains. piMASS was run for 1.1×10^7

iterations with 10^6 as burn-in. GUESS analysis was performed on an HPC cluster computer with a 2.8 GHz Dual-Core Xeon processor and a NVidia Tesla C1060 GPU with 8 Gb of RAM, while piMASS was run on a 3 GHz computer with a 1024 KB cache size Dual-Core AMD Opteron processor and 16 Gb of RAM. “Computational time” is reported in hours (rounded to the nearest integer). “Number of models evaluated” includes the burn-in phase, while “Number of unique model visited” and “Number of model visited before (visiting) top Best Model Visited” are calculated after the burn-in phase. “Average model size” is the average dimension (standard deviation in brackets) of the model recorded in GUESS (from the non-heated chain) and piMASS (every 10 iterations). For piMASS the number of models evaluated corresponds to the number of iterations and is roughly equal to a third of the models evaluated by GUESS.

(PDF)

Table S4 Output obtained from SNPTEST for all elements of the two trees (green and blue colour coded) and TG-HDL-LDL. Horizontal lines separating groups of traits with the same cardinality (singleton, pairs and triplets). The unique set of significant SNPs ($FDR < 0.05$) found by GUESS which predict a group of phenotypes is indicated on the top of the table as well as the associated locus. Based on Ensembl R66 annotation, each locus is classified as: (1) intronic, (2) 3’UTR, (3) downstream, (4) previously associated and (5) tagSNP of previously associated SNP. In the centre of the table the SNPTEST \log_{10} (Bayes Factor) for significant SNPs found associated by GUESS is included (the \log_{10} (BF) is truncated at 20).

(PDF)

Table S5 Results of the empirical FDR procedure. For each element of the two trees centred on the LDL and HDL pathways and TG-HDL-LDL, we report the sample size of the null distribution used in the empirical FDR procedure that we obtained combining the Marginal Posterior Probability of Inclusion (MPPI) for all the artificial groups of phenotypes with the same dimension (each element of the trees was permuted 3 times and the MPPI of all artificial groups of traits with the same dimension, i.e. 5 singleton (+), 6 pairs (++) and 3 triplets (+++), were used to calculate the empirical FDR), the MPPI threshold at an FDR of 5% and the number of significant SNPs associated with each group of phenotypes.

(PDF)

Table S6 Genetic associations for selected phenotypic groups (TG-APOB and TG-LDL-APOB in “Tree I”; TG-HDL in “Tree II” and TH-HDL-LDL) detected by GUESS. Two independent replication datasets were used (a) Copenhagen City Heart Study (CCHS) and (b) Data from an Epidemiological Study on the Insulin Resistance syndrome (DESIR). Each region centred at the identified causal variant in the discovery dataset (rs629301 and rs261333, respectively) and spanning 2 Mb is regressed against the

phenotypic groups previously associated with the variant. Genetic markers with the largest significant BF obtained with SNPTEST in each region are reported in each table as well as their position in the genome. SNP rs261332 is not present in the CardioMeta-chip. Using 381 Caucasian individuals from the 1000 Genomes project, r^2 and D' are 0.582 and 0.979 between rs261332 and rs8034802 in (A) and 0.838 and 0.982 between rs261332 and rs1077834 in (B), respectively.

(PDF)

Table S7 Strength of the genetic association provided by the RBF between genetic variants identified in Table S6 and the branches of the two trees and TG-HDL-LDL. Two independent replication datasets were used (a) Copenhagen City Heart Study (CCHS) and (b) Data from an Epidemiological Study on the Insulin Resistance syndrome (DESIR) (computations of RBF based on the SNPTEST BF). Combinations of phenotypic groups and genetic markers previously found to be most associated in the discovery data set Gutenberg Health Study (GHS) are highlighted in bold. A dashed line indicates that the genetic association is not significant at a 5% FDR in the selected 2 Mb region.

(PDF)

Table S8 Post-processed output obtained from GUESS with different prior specification. For selected elements of the two trees (green and blue colour coded) and TG-HDL-LDL GUESS was run using a conjugate hierarchical independent prior for the genetic effects with a diffuse exponential hyper-prior for the variance of the regression coefficients. Model Posterior Probability (MPP) of the top Best Model Visited (BMV) and the cumulative MPP of the five top BMV are indicated in the first two columns of the table, respectively. The previously identified unique sets of significant SNPs ($FDR < 0.05$) which predict a group of phenotypes is indicated on the top of the table as well as the associated locus. Based on Ensembl R66 annotation, each locus previously identified is classified as: (1) intronic, (2) 3’UTR, (3) downstream, (4) previously associated and (5) a tagSNP of a previously associated SNP. SNP-trait(s) association identified in the BMV by GUESS with the new prior specification are presented in the centre of the table. The BMV for the selected elements of the two trees are as depicted in Table S1.

(PDF)

Author Contributions

Conceived and designed the experiments: LB SR LT FC DT EP. Analyzed the data: LB MCH DIH BL PN LY RC MF. Contributed reagents/materials/analysis tools: MCH DIH LB. Wrote the paper: LB SR LT EP. Gutenberg Health Study (GHS) data generation: SB TZ PSW AS AZ TM. Copenhagen City Heart Study (CCHS) data generation: JD ASB WKH SFN BGN ATH. Data from an Epidemiological Study on the Insulin Resistance syndrome (DESIR) data generation: PF FF.

References

- Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, et al. (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* 41: 677–687.
- Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466: 707–713.
- Brown PJ, Vannucci M, Fearn T (1998) Multivariate Bayesian variable selection and prediction. *J Roy Stat Soc B* 60: 627–641.
- Denison DGT, Holmes CC, Mallick BK, Smith AFM (2002) Bayesian Methods for Nonlinear Classification and Regression. New York: Wiley.
- Monni S, Tadesse MG (2009) A stochastic partitioning method to associate high-dimensional responses and covariates (with discussion). *Bayesian Analysis* 4: 413–436.
- Chipman H, George EI, McCulloch RE (2001) The practical implementation of Bayesian model selection (with discussion). In: Lahiri P, editor. *Model Selection*. Beachwood: IMS.
- Bottolo L, Richardson S (2010) Evolutionary Stochastic Search for Bayesian model exploration. *Bayesian Analysis* 5: 583–618.
- Guan Y, Stephens M (2011) Bayesian variable selection regression for Genome-Wide Association Studies, and other large-scale problems. *Ann Appl Stat* 5: 1780–1815.
- Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39: 906–913.
- Nath R, Pavur R (1985) A new statistic in the one way multivariate analysis of variance. *Comp Stat Data An* 2: 297–315

11. Friedman J, Hastie T, Tibshirani R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33: 1–22.
12. Shen H, Huang JZ (2008) Sparse principal component analysis via regularized low rank matrix approximation. *J Multivariate Anal* 99: 1015–1034.
13. Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26: 445–455.
14. Wilson MA, Iversen ES, Clyde MA, Schmidler SC, Schildkraut JM (2010) Bayesian model search and multilevel inference for SNP association studies. *Ann Appl Stat* 4: 1342–1364.
15. Deukwoo D, Landi MT, Vannucci M, Issaq HJ, Prieto DR, et al. (2011) An efficient stochastic search for Bayesian variable selection with high-dimensional correlated predictors. *Comput Stat & Data Anal* 55: 2807–2818.
16. Zhang W, Zhu J, Schadt EE, Liu JS (2010) A Bayesian partition model for detecting pleiotropic and epistatic eQTL modules. *PLoS Comput Biol* 6: e1000642.
17. Lee A, Yau C, Giles M, Doucet A, Holmes C (2010) On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *J Comput Graph Statist* 19: 769–789.
18. Wakefield J (2009) Bayes factors for genome-wide association studies: comparison with p-values. *Genet Epidemiol* 33: 79–86.
19. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. *Nat Rev Gen* 10: 681–690.
20. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *J R Statist Soc B* 58: 267–288.
21. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 96: 1348–1360.
22. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Statist Soc B* 67: 301–320.
23. Zou H (2006) The adaptive Lasso and its oracle properties. *J Am Stat Assoc* 101: 1418–1429.
24. Tibshirani R, Saunders M, Rosset S, Ji Z, Knight K (2005) Sparsity and smoothness via the fused Lasso. *J R Statist Soc B* 67: 91–108.
25. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by Lasso penalized logistic regression. *Bioinformatics* 25: 714–721.
26. Leng C, Lin Y, Wahba G (2006) A note on the Lasso and related procedures in model selection. *Statist Sin* 16: 1273–1284.
27. Jin L (2011) Penalized methods in genome-wide association studies: University of Iowa. Available: <http://ir.uiowa.edu/etd/1242/> Accessed 30 June 2013.
28. Meinshausen N, Bühlmann P (2010) Stability selection (with discussion). *J R Statist Soc B* 72: 417–473.
29. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, et al. (2010) Genetics and beyond: the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* 5: e10693.
30. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, et al. (2010) From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* 466: 714–719.
31. Kjolby M, Andersen OM, Breiderhoff T, Fjorback AW, Pedersen KM, et al. (2010) *SORT1*, encoded by the cardiovascular risk locus 1p133, is a regulator of hepatic lipoprotein export. *Cell Metab* 12: 213–223.
32. Annema W, Tietge UJ (2011) Role of hepatic lipase and endothelial lipase in high-density lipoprotein-mediated reverse cholesterol transport. *Curr Atheroscler Rep* 13: 257–265.
33. Petersen AK, Stark K, Musameh MD, Nelson CP, Römisch-Margl W, et al. (2012) Genetic associations with lipoprotein subfractions provide information on their biological nature. *Hum Mol Genet* 21: 1433–1443.
34. McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet* 17: R156–165.
35. Braun TR, Been LF, Singhal A, Worsham J, Ralhan S, et al. (2012) A replication study of GWAS-derived lipid genes in asian indians: the chromosomal region 11q233 harbors loci contributing to triglycerides. *PLoS ONE* 7: e37056.
36. Nordestgaard BG, Benn M, Schnohr P, Tybjaerg-Hansen A (2007) Nonfasting triglycerides and risk of myocardial infarction, ischemic heart disease, and death in men and women. *JAMA* 298: 299–308.
37. Nordestgaard BG, Palmer TM, Benn M, Zacho J, Tybjaerg-Hansen A, et al. (2012) The effect of elevated body mass index on ischemic heart disease risk: causal estimates from a Mendelian randomisation approach. *PLOS Med* 9: e1001212.
38. Balkau B (1996) An epidemiologic survey from a network of French Health Examination Centres, (D.E.S.I.R.): epidemiologic data on the insulin resistance syndrome. *Rev Epidemiol Sante Publique* 4: 373–375.
39. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, et al. (2012) The Metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet* 8: e1002793.
40. McVean, et al (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56–65.
41. Kim S, Xing EP (2009) Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet* 5: e1000587.
42. Petretto E, Bottolo L, Langley SR, Heinig M, McDermott-Roe MC, et al. (2010) New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Comput Biol* 6: e1000737.
43. Storey JD (2002) A direct approach to false discovery rates. *J R Statist Soc B* 63: 479–98.
44. Yang J, Ferreira T, Morris AP, Medland SE, et al. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 44: 369–36.
45. Judd CM, McClelland GH, Ryan CS (2009) *Data Analysis: A Model Comparison Approach*. London: Routledge.
46. Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial (with discussion). *Stat Sci* 14: 382–401.
47. Hartung J, Knapp G, Sinha BK (2008) *Bayesian Meta-Analysis, in Statistical Meta-Analysis with Applications*. New York: John Wiley & Sons, Inc.
48. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78: 629–644.
49. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, et al. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 74: 106–120.
50. Bottolo L, Chadeau-Hyam M, Hastie DI, Langley SR, Petretto E, et al. (2011) ESS++: a C++ objected-oriented algorithm for Bayesian stochastic search model exploration. *Bioinformatics* 27: 587–588.
51. Barbieri MM, Berger JO (2004) Optimal predictive model selection. *Ann Stat* 33: 870–897.
52. Chen W, Ghosh D, Trivellore E, Raghunathan TE, Sargent DJ (2009) Bayesian Variable Selection with joint modelling of categorical and survival outcomes: an application to individualizing chemotherapy treatment in advanced colorectal cancer. *Biometrics* 65: 1030–1040.
53. Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* 3: e114.
54. Kass RE, Raftery AE (1995) Bayes Factors. *J Am Stat Assoc* 90: 773–79.
55. Lê Cao K-A, Rossouw D, Robert-Granić C, Besse P (2008) A sparse PLS for variable selection when integrating Omics data. *Stat App Gen Mol Biol* 7 article 35.
56. Lê Cao K-A, Martin PGP, Robert-Granić C, Besse P (2009) Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* 10:34.