Sherlock: Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS

Xin He,[1,2] Chris K. Fuller,[1] Yi Song,[1] Qingying Meng,[3] Bin Zhang,[4] Xia Yang,[3] and Hao Li[1,*]

Genetic mapping of complex diseases to date depends on variations inside or close to the genes that perturb their activities. A strong body of evidence suggests that changes in gene expression play a key role in complex diseases and that numerous loci perturb gene expression in *trans*. The information in *trans* variants, however, has largely been ignored in the current analysis paradigm. Here we present a statistical framework for genetic mapping by utilizing collective information in both *cis* and *trans* variants. We reason that for a disease-associated gene, any genetic variation that perturbs its expression is also likely to influence the disease risk. Thus, the expression quantitative trait loci (eQTL) of the gene, which constitute a unique "genetic signature," should overlap significantly with the set of loci associated with the disease. We translate this idea into a computational algorithm (named Sherlock) to search for gene-disease associations from GWASs, taking advantage of independent eQTL data. Application of this strategy to Crohn disease and type 2 diabetes predicts a number of genes with possible disease roles, including several predictions supported by solid experimental evidence. Importantly, predicted genes are often implicated by multiple *trans* eQTL with moderate associations. These genes are far from any GWAS association signals and thus cannot be identified from the GWAS alone. Our approach allows analysis of association data from a new perspective and is applicable to any complex phenotype. It is readily generalizable to molecular traits other than gene expression, such as metabolites, noncoding RNAs, and epigenetic modifications.

Introduction

Recent application of genome-wide association studies (GWASs) to complex human diseases led to the discovery that the majority of disease-associated variants (estimated to be as high as 88%) are located in noncoding sequences, potentially affecting gene expression rather than protein function.[1,2] Because of the complexity of gene regulation, the expression of a gene can be modulated by mutations in *cis* (proximal to the gene) and/or in *trans* (distal to the gene or on different chromosomes, such as upstream transcription/chromatin factors, distal regulatory elements, etc.).[3,4] As a result of a large mutational target size (primarily because of mutations in *trans*) and the buffering of gene regulatory systems that helps tolerate expression changes, genetic variants altering expression levels are common in populations.[5] Indeed, many studies of expression quantitative trait loci (eQTL) demonstrate that the expression of most genes is influenced by multiple loci, most of which act in *trans*.[6–9] Despite their individually small effect sizes, *trans* eQTL are collectively important for variation of gene expression and by some estimates account for a larger proportion of the heritability of gene expression than do *cis* eQTL.[8,10]

Because of their prevalence in the human population, expression variations, especially those in *trans,* provide systematic perturbations of the gene regulatory networks underlying various complex phenotypes, and as such might reveal important information about the genetic basis of these phenotypes. Thus there is a pressing need to develop a general framework to mine the collective information in both *cis*- and *trans*-expression QTL in the context of association studies. So far, information from *trans* variations has largely been ignored because only *cis* variants can be assigned to their target genes based on proximity by using the GWAS data alone. The growing collection of eQTL data for various human tissues makes it possible to associate *trans* variants with target genes.[8] Although previous studies demonstrated the utility of eQTL data for aiding the analysis of association studies,[9,11] most of these used only *cis* eQTL located close to the genes. This reflects some fundamental difficulties of utilizing information in *trans*. Because *trans* eQTL are usually much weaker than those in *cis*, the statistical signal of an individual *trans* eQTL is difficult to detect—it may fall far below the genome-wide threshold.[9] Another major challenge is the pleiotropic effect of *trans* variation. A gene's *trans* perturbation may come from the mutation of a regulatory molecule, but this mutation may also affect multiple other genes.

Here we present a general strategy to infer genes whose perturbations contribute to the etiology of complex diseases by tapping into statistical information provided by both *cis* and *trans* variations affecting gene expression. Although individual variants are often weak and not particularly informative, the overall pattern of expression variants of a gene can provide a strong statistical signal. A unique aspect of this strategy is that because we utilize *trans* variants far from target genes, it is possible to identify important genes distal to any GWAS association signals and thus impossible to detect with GWAS alone.

[1]Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, CA 94143, USA; [2]Lane Center of Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA; [3]Department of Integrative Biology and Physiology, University of California at Los Angeles, Los Angeles, CA 90095, USA; [4]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA
*Correspondence: haoli@genome.ucsf.edu

http://dx.doi.org/10.1016/j.ajhg.2013.03.022.

To illustrate the basic concept, we consider a gene whose expression level may influence the risk of a complex disease. There may be multiple variants in the genome that affect the expression of this gene in a disease-related tissue (these are expression SNPs, or eSNPs; technically eSNPs are only in linkage disequilibrium with the causal variants, though we do not make the distinction here). A change of genotype at any of these eSNPs will lead to a change of expression level, which could in turn alter the disease risk (Figure 1A). Therefore, many of these eSNPs are likely to be associated with the disease as well. In general, each gene has a different set of eSNPs across the genome, with different effect sizes, reflecting the unique regulatory program governing the expression of this gene. The characteristic pattern of all genetic associations with the expression of a gene thus constitutes a unique "genetic signature" of this gene (Figure 1B, top). Similarly, we can define the genetic signature of a phenotype as the statistical association of all the loci in the genome to the phenotype. If the expression level of a gene influences the disease risk, the genetic signature of this gene should match, at least partially, the genetic signature of the disease (Figure 1B, bottom).

Our analysis identifies potential gene-disease association by matching the genetic signature of a gene to that of the disease, using the GWAS data of the disease and the eQTL data of a related tissue. It is worth emphasizing that we do not directly test the statistical relationships between genotypes, gene expression, and phenotypes, as done by earlier methods.[12] Thus, the eQTL and GWAS data do not have to come from the same subjects. Instead, the eQTL data provide information of the genetic signatures of many genes, and then in an independent GWAS of some phenotype, the match of a gene's signature with the GWAS would suggest that the gene plays a role in that phenotype. This is much like how a detective works (hence the name of our algorithm, "Sherlock"): he compares the fingerprint from a crime scene (like our GWAS association data) against a database of fingerprints (like our eQTL data) to determine the real culprit (the genes whose expression levels influence the disease risk).

We implement this idea of genetic signature matching by using a Bayesian statistical framework. Instead of applying a stringent cutoff, we utilize both strong and moderate SNPs in the eQTL and GWAS data. The statistical model allows us to access information in the moderate SNPs without introducing many false signals. Application to two well-studied diseases shows the promise of our approach. We predicted ten genes associated with Crohn disease (MIM 266600), six of which are highly plausible based on literature evidence. With an independent GWAS data set, all but one gene were replicated. Among our four predicted genes associated with type 2 diabetes (T2D [MIM 125853]), three are supported by experimental evidence from literature, and the other is a promising candidate based on a combined analysis of multiple genomic data sets.

In summary, our approach allows the analysis of association studies from a different perspective and, as we demonstrate, enables the discovery of genes and pathways missed by the traditional GWAS analysis. We have constructed a web-based resource to facilitate the application of our method. We collected eQTL data sets from multiple human tissues and provided the software to search for gene-disease associations with the disease GWAS as a query (only p values are needed). With the increasing collection of eQTL data (as well as QTL of other molecular phenotypes), we expect that these resources/tools will become an important platform for interpreting the results from association studies.

Material and Methods

Statistical Model

Given a gene and the disease of interest, our method tests whether the expression change of the gene has any effect on the risk of the disease, using the information of N putative eSNPs of the gene (passing some low significance threshold in the eQTL data). We define binary indicator variables U_i and V_i to represent whether the i^{th} SNP is associated with the expression and the disease trait, respectively, and a binary indicator variable Z to represent whether the gene is associated with the disease (Figure 1A). Our data consist of the p values of SNPs relative to the gene expression trait (eQTL profile), denoted as vector $\boldsymbol{x}$, and the p values of the SNPs relative to the phenotypic trait (GWAS profile), denoted as $\boldsymbol{y}$. Although U_i and V_i are not observed, they are related to x_i and y_i: when $x_i(y_i)$ is small (i.e., the SNP is highly significant), it is likely that $U_i(V_i)$ is true. We use the data $\boldsymbol{x}$ and $\boldsymbol{y}$ to test the hypothesis H_0 (i.e., $Z = 0$) that the gene is not associated with the disease versus the alternative hypothesis (H_1). The dependencies of the statistical variables are shown in Figure 1D. We describe some intuitions of our model. Under H_0, the gene is irrelevant to the phenotype, and thus U_i and V_i, are independent in a statistical sense (Figure 1D, top). Under H_1, a true eSNP of the gene is expected to be associated with the phenotype (Figure 1A). In other words, whenever U_i is true, V_i should also be true under H_1 (Figure 1D, bottom).

Our inference is based on the posterior ratio, defined as:

$$\frac{P(Z=1\,|\,x,y)}{P(Z=0\,|\,x,y)}=\frac{P(Z=1)}{P(Z=0)}\times\frac{P(x,y\,|\,Z=1)}{P(x,y\,|\,Z=0)}. \qquad \text{(Equation 1)}$$

The first term in the right side is the ratio of the prior probabilities of H_1 and H_0. Typically, the prior ratio is small because only a small fraction of all genes are expected to be associated with a disease. The second term is the ratio of model evidence (i.e., the probability of data under H_1 or H_0) called the Bayes factor (BF). The BF is similar to the familiar likelihood ratio test and does not depend on the prior probabilities. We will focus on computing the BF for the inference task.

We assume all SNPs are unlinked (see below for LD blocks). The BF of the gene expression trait is the product of the BF of all SNPs:

$$B=\prod_i B_i=\prod_i\frac{P(x_i,y_i\,|\,Z=1)}{P(x_i,y_i\,|\,Z=0)}. \qquad \text{(Equation 2)}$$

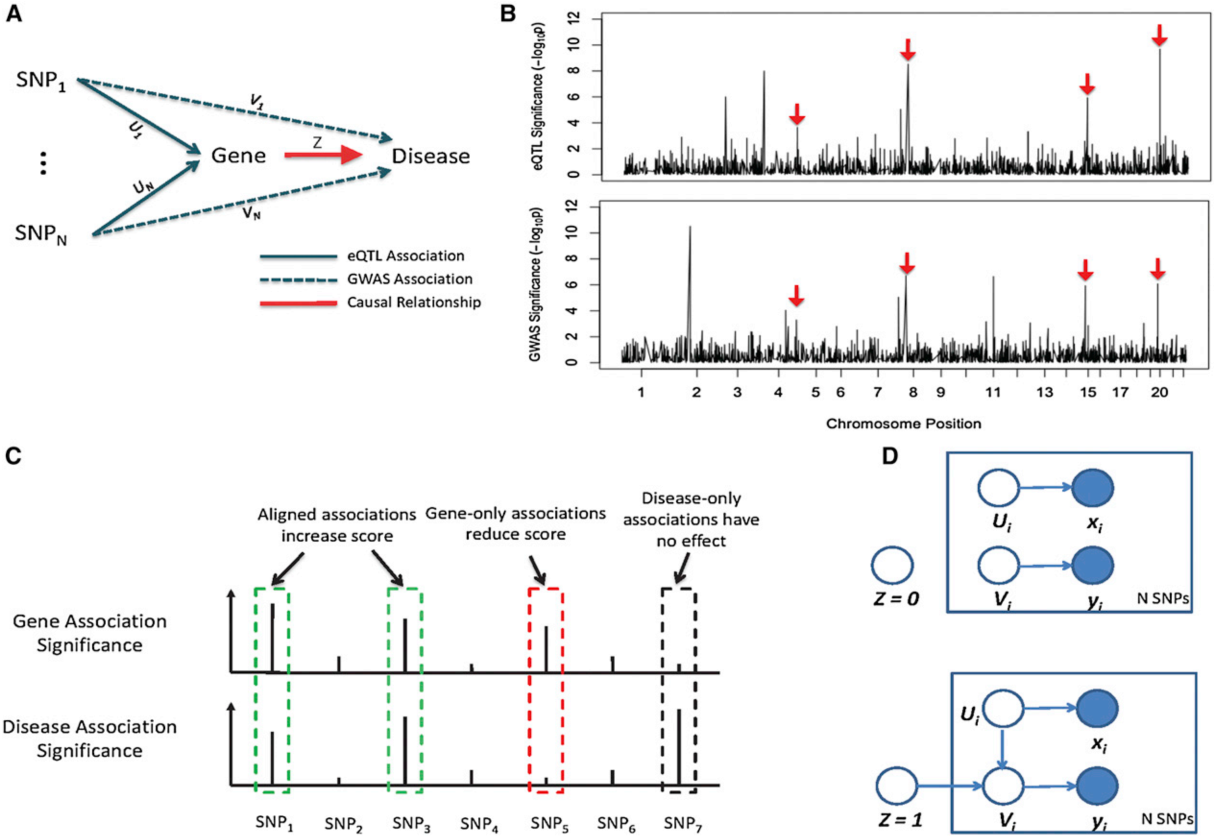
A
SNP1
SNPN
U1
UN
V1
VN
Gene
Z
Disease
eQTL Association
GWAS Association
Causal Relationship
B
eQTL Significance (−log10P)
GWAS Significance (−log10P)
0
2
4
6
8
10
12
1
2
3
4
5
6
7
8
9
11
13
15
17
20
Chromosome Position
C
Aligned associations increase score
Gene-only associations reduce score
Disease-only associations have no effect
Gene Association Significance
Disease Association Significance
SNP1
SNP2
SNP3
SNP4
SNP5
SNP6
SNP7
D
Z = 0
Ui
xi
Vi
yi
N SNPs
Z = 1
Ui
xi
Vi
yi
N SNPs

Figure 1. The Sherlock Algorithm: Matching Genetic Signatures of Gene Expression Traits to that of the Disease to Identify Gene-Disease Associations

(A) Perturbation of the expression level of a disease-associated gene at any of its eQTL changes the disease risk, and thus the eQTL tend to be associated with the complex disease as well (the dashed lines). The eQTL associations may contain false positives, so we use binary indicator variables, U, to represent the true SNP-gene expression relationship; similarly we use indicator variables, V, for the SNP-disease relationship. Z is a binary variable indicating whether the expression trait influences the disease risk.

(B) Hypothetic genome-wide association plots of the causal expression trait (top) and a complex disease (bottom). The genetic signature of the gene expression trait partially overlaps with that of the disease. Red arrows indicate the matched loci.

(C) Alignment of genetic signatures of a gene expression trait and the phenotype. Three different scenarios are shown, represented by the green, red, and black boxes.

(D) The probabilistic model representing the dependency of the variables. The semantics of the variables U, V, and Z are shown in (A). When $Z = 0$, U and V are independent (top). When $Z = 1$, V depends on both Z and U; if $U = 1$, then V is also likely to be 1 (bottom). The association statistics of a SNP with respect to the gene expression trait and the disease (x and y) depend on the hidden variables U and V. Shaded and open circles indicate observed and latent variables, respectively.

The likelihood function at each SNP at a given Z is computed by summing over the hidden variables U_i and V_i:

$$P(x_i, y_i \mid Z) = \sum_{U_i,V_i} P(U_i)P(V_i \mid Z, U_i)P(x_i \mid U_i)P(y_i \mid V_i). \qquad \text{(Equation 3)}$$

U_i is a Bernoulli random variable with the success probability α (the prior probability of a SNP being associated with the expression trait). The information about the gene-disease relationship is encoded in the conditional probability $P(V_i \mid Z, U_i)$. When $Z = 0$ or when $Z = 1$ and $U_i = 0$ (a false eSNP), V_i is a Bernoulli random variable with the success probability β (the prior probability of a SNP being associated the phenotype). When $Z = 1$ and $U_i = 1$, according to our discussion before, V_i should also be 1. The probability terms $P(x_i \mid U_i)$ and $P(x_i \mid V_i)$ reflect the distribution of p values under the null or alternative hypothesis, and we derive these distributions in Appendix A.

The Bayes factor defined in Equation 2 can be expressed in terms of the following variables:

$$B_{i,x} = \frac{P(x_i \mid U_i = 1)}{P(x_i \mid U_i = 0)} \quad B_{i,y} = \frac{P(y_i \mid V_i = 1)}{P(y_i \mid V_i = 0)}. \qquad \text{(Equation 4)}$$

These are the Bayes factors measuring the association of the i^{th} SNP with the expression and the phenotypic trait, respectively. We show, in Appendix B, that the BF of the i^{th} SNP in Equation 2 is given by

$$B_i = \frac{1-\alpha}{1-\alpha+\alpha B_{i,x}} + \frac{\alpha B_{i,x}}{1-\alpha+\alpha B_{i,x}} \frac{B_{i,y}}{1-\beta+\beta B_{i,y}}. \qquad \text{(Equation 5)}$$

Thus the Bayes factor of the gene being tested depends only on the parameters α, β, and the SNP-level Bayes factors. If Bayesian inference has been performed in both the eQTL and GWAS analysis, it is straightforward to combine the resulting BFs to obtain the BF for the gene.

Dealing with Linkage Disequilibrium

For multiple adjacent SNPs, we use a block-level BF in Equation 2. According to Duan et al.,[13] a block is defined as a region of the genome containing one or more eSNPs associated with the same gene and having a between-eSNP interval of <500 Kb. Although this criterion appears somewhat arbitrary, it does not have a large impact on our analysis primarily because by defining eSNPs with a certain threshold ($p < 10^{-5}$), most eSNPs naturally fall into blocks. To combine the BFs of individual SNPs in a block, we follow Servin and Stephens:[14] the block-level BF is the mean of the BFs of all SNPs in that block. It has been shown that this simple averaging is a reasonable way of dealing with dependent SNPs in a region.[14]

Model Analysis

To gain some intuition into our method, we show that it leads to a scoring scheme that is similar to sequence alignment but with an inherent asymmetry (Figure 1C).[15] According to Equation 2, the total score (logarithm of BF, or LBF) of a gene is the sum of LBF of each SNP, written as $S_i = \log B_i$. We discuss the value of S_i for different types of SNPs. We first note that in Equation 5, both α and β are small numbers (prior probabilities, typically less than 0.01). For a SNP not associated with the gene expression trait, $B_{i,x} < 1$, thus $\alpha B_{i,x} \ll 1$, according to Equation 5, $S_i \approx 0$. These SNPs are not informative, regardless of their association status with the phenotype. For the informative SNPs, we consider only very strong eSNPs for simplicity of analysis, $\alpha B_{i,x} \gg 1$, and we have this approximation:

$$S_i = \log B_i \approx \log \frac{B_{i,y}}{1 + \beta B_{i,y}}. \qquad \text{(Equation 6)}$$

We analyze three possible cases depending on the strength of association of a SNP with the phenotype. (1) The SNP is strongly associated with the phenotype $\beta B_{i,y} \gg 1$, and thus we have $S_i \approx \log(1/\beta)$. This is a positive number and represents the reward when the signatures of the two traits match at this SNP. (2) The SNP is moderately associated with the phenotype $B_{i,y} > 1$ but $\beta B_{i,y} \ll 1$, and thus we have $S_i \approx \log B_{i,y} > 0$, representing the positive but smaller contribution of the SNP. (3) The SNP is not associated with the phenotype $B_{i,y} < 1$, and thus we have the approximation $S_i \approx \log B_{i,y} < 0$. The negative score represents the evidence against the gene.

This analysis allows us to understand some properties of the model. First, the score of any SNP is always bounded by $\log(1/\beta)$, even if the SNP reaches extremely low p values. Thus the LBF of a gene is generally not dominated by a single SNP (except for genes with relatively few eSNPs). Another property of the model is that the score of a SNP moderately associated with the phenotype ($S_i \approx \log B_{i,y}$) is not sensitive to the parameter β. Thus we see that β determines the relative contribution of strong versus weak SNPs: a smaller value of β favors a single strong SNP over multiple weaker ones. In general, we believe that genes supported by multiple SNPs are more interesting than those supported by single strong SNPs, which are probably due to the pleiotropic effects of SNPs (SNPs independently associated with gene expression and the phenotype). This suggests that we should use a relatively high value of β. Meanwhile, this would lead to a conservative estimate of the Bayes factor (because the maximum LBF per SNP is $\log(1/\beta)$).

Statistical Significance of Bayes Factors

We use simulation to compute the p values of the BFs, a procedure known as Bayes/non-Bayes compromise.[14] To generate the null distribution of the BFs, we fix the eQTL profiles (the p values of putative eSNPs) of all the genes and randomize the GWAS data K times (see below for the details of randomization). The resulting BFs of all genes over all the K-simulated GWAS data sets then form the genome-wide null distribution. The p value of each gene is estimated from the ranking of its BF in this null distribution. We used $K = 50$ in our experiments.

To create a randomized GWAS data set, we follow the procedure described in Liu et al.[16] Note that we need to generate randomized GWAS data only for all the putative eSNPs across all genes, because other SNPs will not enter the BF calculation. For each gene, we first divide all its putative eSNPs into blocks, then the p values of each block are sampled independently: within each block, a multivariate normal (MVN) random vector is sampled with the covariance matrix matching the LD structure of the block, and the vector is then converted to p values of SNPs. It has been shown that this MVN-based approach is a very good approximation of the full permutation procedure (random swapping of cases and controls).[16]

We choose not to perform permutation of eQTL data because our statistical procedure depends on the alignment of the eQTL and GWAS profiles, which are relative to each other. Thus permuting GWAS data should be equivalent to permuting eQTL data. In practice, permuting eQTL data is more difficult to implement, because there are correlated structures in eQTL data that are difficult to account for and the genotype data of eQTL are generally not available.

Choice of Parameters

The parameters α and β specify the prior probabilities of a SNP being associated with an expression and a phenotypic trait, respectively. We further distinguish between *cis* and *trans* eSNPs: α should be higher for *cis* eSNPs (within 1 Mb of the gene) than for *trans* eSNPs. As per the guidelines in literature,[17] we chose these values: $\alpha = 1.0 \times 10^{-3}$ (*cis*) and 5.0×10^{-5} (*trans*), $\beta = 1.0 \times 10^{-3}$. We provide some intuitive explanation of these parameters. If we assume there are a total of one million SNPs, the number of SNPs close to a gene (within 1 Mb of the coding sequence) is roughly $1{,}000{,}000/3{,}000{,}000{,}000 \times 2{,}000{,}000 = 1{,}000$ (only order-of-magnitude estimate is made here). Assuming a gene has one *cis* eSNP, the prior probability α for SNPs in *cis* is $1/1{,}000 = 10^{-3}$. For *trans* eSNP, we assume that a gene may have a relatively large number of eSNPs across the genome, say 50, then the prior probability for SNPs in *trans* is $50/1{,}000{,}000 = 5.0 \times 10^{-5}$. Our selection of the value of β is based on the following: (1) complex traits are known to be associated with hundreds of loci,[18] and (2) according to the discussion in the section "Model Analysis," a somewhat high value of β is preferred for our model. We chose $\beta = 10^{-3}$ in our experiments.

For a SNP associated with a trait, we assume that its effect size follows a prior normal distribution $N(0, \sigma_a^2)$. The default value of the prior variance parameter is 0.5 for expression traits and 0.2 for phenotypic traits, based on earlier studies.[14,17] We note that these parameters are not necessary if the Bayes factors of SNP associations are available from eQTL and GWAS analysis, according to Equation 5. In addition to these prior parameters, the method also requires the disease prevalence (5.0×10^{-4} for Crohn disease and 0.1 for type 2 diabetes) and allele frequencies for each SNP (from HapMap).

Computational Experiments

The eQTL and GWAS data were downloaded from their respective sources. Because only putative eSNPs are informative, we applied a weak cutoff to eQTL based on p values: $p < 10^{-5}$ for *trans* associations and $p < 10^{-4}$ for *cis* associations. Each gene with at least one putative eSNP is scored by our program.

In the replication experiment for Crohn disease, we took the meta-analysis data from IBD Genetics Consortium.[19] Because this data set includes some data we used for predicting disease genes, we removed it by using the weighted subtraction algorithm in Zhong et al.,[20] and this gave about 2,000 cases and 7,000 controls.

Coexpression Analysis of *PURB*

The gene expression data were taken from several studies in humans[21–23] and various mouse crosses.[12,24,25] The weighted coexpression networks were constructed with previously described methods to derive subnetworks or modules containing highly coregulated genes from each tissue.[26] We retrieved all network modules that contain *PURB* from all coexpression networks and then extracted all genes that share module membership with *PURB* as *PURB*-coexpressed genes. These genes are then ranked by how often they appear with *PURB* in the same modules.

Results

Genetic Signature Matching Algorithm: Sherlock

Our algorithm takes as input the association results from independent eQTL and GWAS experiments and aims to identify genes whose expression levels may influence the disease risk. It uses summary statistics (p values of the associations), but does not require the genotype and phenotype data. As discussed previously, it assumes that such genes will have multiple, coincident loci with elevated significance in both the GWAS and eQTL data sets (Figures 1A and 1B). The algorithm scores the significance of the overlap by the statistical model described in Material and Methods.

To see why significant overlap supported by multiple loci may imply causality, we consider two alternative scenarios.[12] In the first scenario, a SNP may affect the disease and also happen to affect the expression of a gene that has nothing to do with the disease. However, the likelihood that multiple disease-associated SNPs happen to affect the expression of the same gene by chance is very small. In the second scenario, the disease may compromise the gene expression patterns of relevant cells, causing some expression traits to share loci with the disease. This scenario is also unlikely because the eQTL mapping used for our analysis is usually performed in individuals unrelated to the GWAS phenotype, and thus the disease loci generally should not affect gene expression in these samples.

Intuitively, our algorithm performs an alignment of the genetic signature of one phenotypic trait against an expression trait. For any SNP, there are potentially three scenarios, depending on the association of the SNP with the two traits (Figure 1C). Any SNP that appears to be associated with both traits contributes a positive score to the gene (Figure 1A). Any SNP associated with the expression trait but not the disease is evidence against the role of the gene, contributing a negative score. The third type of SNPs are associated with the disease but not the expression of the gene being tested. Because they are not informative for this gene at the expression level, they will not contribute to the score of the gene.

We compute the Bayes factor (or its logarithm, LBF) for each gene (see Material and Methods), which evaluates evidence supporting that the gene is associated versus not associated with the disease. Specifically, we first compute the LBF score of each putative eSNP of the gene being tested, which represents how strongly the SNP supports a functional role of this gene (see Material and Methods). The LBF score of a SNP is analogous to the nucleotide matching scores in sequence alignment, with the sign of the score corresponding to the three situations described above (also see Material and Methods, "Model Analysis"). For the cases where multiple SNPs in LD are associated with a gene, we define a block-level LBF and treat them as if they were a single SNP (see Material and Methods, "Dealing with Linkage Disequilibrium"). Herein, the set of all eSNPs of a gene are represented by a set of independent eSNPs, each of which may actually represent a block of adjacent SNPs. The total LBF score of a gene is the sum of LBFs of all SNPs. The value of the LBF score of a gene reflects the strength of evidence: for example, a LBF of 2.3 means that the posterior probability of the gene being associated with the disease is $exp(2.3) = 10$ times more likely than the opposite hypothesis assuming that the two hypotheses are equally likely a priori. In practice, a gene is a priori far more likely to be unrelated to the disease, so we generally demand a high LBF (e.g., at least 4.0).

Although BFs can be directly used for inference under a purely Bayesian framework,[27] we take an approach, known as Bayes/non-Bayes compromise, to compute the p values of the BFs based on simulations. This has the advantage that the results are less sensitive to the prior assumptions.[14] Ideally, the simulation procedure would permute the case/control labels of the subjects in the GWAS data to obtain the null distribution of the BFs (we discuss in Material and Methods why permutation is applied only to the GWAS and not to the eQTL data). In practice, this requires genotype data (which is generally not publicly available) and is computationally intensive. We use an approximation scheme to generate randomized GWAS data by modeling the LD structure in the genome with multivariate normal (MVN) distributions.[16] We correct for multiple hypothesis testing by using the standard Benjamini-Hochberg procedure.[28]

It is important to emphasize the use of SNPs that fail to reach traditional thresholds for genome-wide statistical significance. For both gene expression and phenotypic traits, recent studies suggest that a large number of loci may have small effects and fall below the statistical threshold.[8,18] Rather than excluding modest GWAS and eQTL associations up front, our statistical method takes them into account and relies on the unlikely occurrence of chance overlap to assign strong significance at the gene level. For example, consider a hypothetical gene

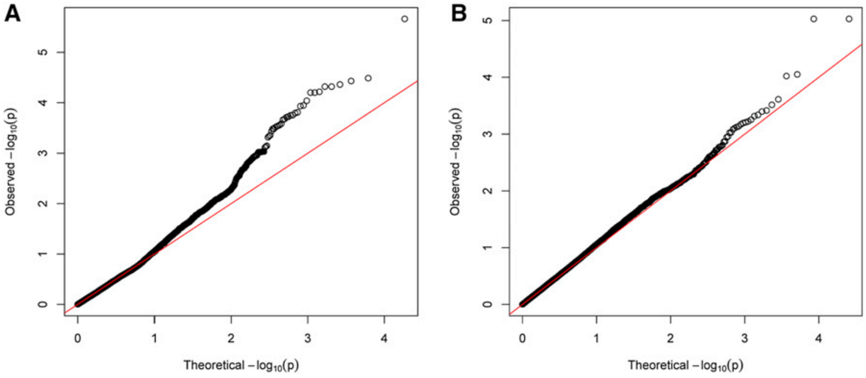

Figure 2. Quantile-Quantile Plots of the p Values at the Log. Scale of All Genes Calculated by Sherlock
(A) Analysis of GWAS data of Crohn disease with the eQTL of lymphoblast B cells.
(B) Analysis of T2D GWAS data with the liver eQTL.

with four independent eSNPs, each having a p value of 0.001 in the GWAS of a disease. Individually, these SNPs are far from achieving genome-wide significance. However, the chance that all four eSNPs achieve such p values in GWAS is extremely low: $0.001^4 = 10^{-12}$. Hence, by combining multiple weak signals, it is possible to achieve strong statistical confidence.

Overview of Experiments

In each experiment, a GWAS data set (summary statistics or p values) of some trait is provided and the program runs the analysis against eQTL data of a relevant tissue in unrelated subjects. For all experiments, the program uses a default set of parameters. For each gene, a LBF score and its p value based on simulation are computed. The distribution of the p values of all genes is analyzed to determine the false discovery rate (FDR) of the predictions. For both Crohn disease and type 2 diabetes, we reported statistically significant findings. We further assessed the predictions by a combination of literature search, replication with independent GWAS, and analysis of additional genomic data sets.

Application to Crohn Disease

Our first experiment is on a GWAS meta-analysis of Crohn disease (3,230 cases and 4,829 controls)[29] uses the eQTL data from lymphoblast B cells (LBL).[13] Crohn disease is an autoimmune disease of the intestines, one of two major forms of inflammatory bowel disease (IBD). The LBL eQTL (mainly *cis* eQTL) has been widely used to support the analysis of GWAS data of immune-related diseases including Crohn disease.[9,11,30] The original GWAS data suffers from modest population stratification ($\lambda = 1.17$), so we applied the genomic control procedure before further analysis.[31] With only *cis* eSNPs, we identified a single SNP, close to *ORMDL3* (MIM 610075), that passes the respective thresholds in both GWAS and eQTL, consistent with the previous findings.[29] In contrast, Sherlock identified a number of significant genes: the Q-Q plot of our predictions for all genes shows a clear excess of genes with small p values (Figure 2A). To rule out p value inflation, we created randomized GWAS data by randomly assigning individuals from the 1000 Genomes Project to “case” and “control” groups.[32] We then performed the same analysis on this randomized data with the real LBL eQTL data set. The p values from this experiment closely follow the uniform distribution, suggesting that our test is well calibrated (Figure S1 available online).

At $p < 10^{-4}$ (the FDR at this p value threshold is 0.09), we predicted ten genes (Table 1). Three of the ten, *PTGER4* (MIM 601586), *ORMDL3*, and *SLC22A5* (MIM 603377), were reported in the original GWAS paper and supported by additional studies.[29,33] Below we discuss the remaining genes and additional analyses. Except for *UBE2L3* (MIM 603721), which is supported by a combination of *cis* and *trans* eSNPs, all the remaining genes are supported only by *trans*-acting SNPs (Table 2, full results in Table S1). Importantly, most of these supporting SNPs have moderate p values in both eQTL and GWAS data and therefore are generally ignored in the traditional GWAS analyses.

Notably, the genetic and experimental evidence strongly suggest possible roles for *UBE2L3* and *EFS* (MIM 609906) in Crohn disease. *UBE2L3* was associated with several autoimmune diseases, including systemic lupus erythematosus (MIM 152700), celiac disease (MIM 212750), and rheumatoid arthritis (MIM 180300).[34] Because autoimmune diseases often share genetic risk loci,[35] the combined evidence supports the role of *UBE2L3* in Crohn disease. *EFS* has not been associated with autoimmune diseases in previous GWASs. However, in mouse studies, *Efs* overexpression was shown to inhibit T lymphocyte development[36] and *Efs* knockout mice exhibit exaggerated T-cell-mediated immune responses.[37] Remarkably, the knockout mice develop tissue-specific inflammatory lesions in their small intestine, a pattern very similar to Crohn disease.[37] Additionally, the target of *EFS*, the kinase *FYN* (MIM 137025), is marginally significant ($p = 9.6 \times 10^{-4}$, FDR < 0.25). Together, this evidence suggests that *EFS* is a strong candidate gene for Crohn disease.

Two other genes in our list have well-established immune functions and are possibly related to autoimmune diseases and IBD. *IK* (MIM 600549) encodes a cytokine that downregulates class II MHC antigen whose aberrant expression has been associated with autoimmune diseases.[38] *LYNX1* (MIM 606110) encodes a signaling peptide of nicotinic acetylcholine receptors (nAChR).[39] The non-neuronal cholinergic system plays an important function in regulating the development and activation of T and B

Table 1. Top Predicted Genes for Crohn Disease

Gene	LBF	p Value	GWAS Hit	Supporting Evidence	p Value (Rep.)
UBE2L3	7.78	2.2×10^{-6}	no	associated with celiac disease, rheumatoid arthritis, and lupus[34]	4.2×10^{-6}
ORMDL3	5.72	3.3×10^{-5}	yes	associated with Crohn disease and ulcerative colitis[29,33]	1.0×10^{-6}
PTGER4	5.57	3.7×10^{-5}	yes	associated with Crohn disease[33]	3.4×10^{-5}
IK	5.43	4.3×10^{-5}	no	regulates class II MHC antigen, which is associated with autoimmune diseases[38]	0.75
LYNX1	5.34	4.8×10^{-5}	no	agonist of nonneuronal nAchR pathway, which may be important for IBD[39–41]	0.014
NUDT4	5.32	4.8×10^{-5}	no	NA	3.4×10^{-5}
EFS	5.23	6.1×10^{-5}	no	knockout mice exhibit symptoms similar to Crohn disease[36,37]	0.039
FAM96A	5.19	6.3×10^{-5}	no	NA	4.2×10^{-6}
SLC22A5	5.17	6.3×10^{-5}	yes	associated with Crohn disease[29,33]	3.8×10^{-5}
ANAPC2	4.87	9.2×10^{-5}	no	NA	0.0006

The LBF column is the logarithm of the Bayes Factor for the genes. The p values refer to the p values of LBFs, calculated from simulations. The "GWAS hit" column shows whether a gene has been implicated as a candidate gene in previous GWASs. The last column shows the p value of the genes using an independent GWAS data set of Crohn disease (replication). See the text for details of the supporting evidence.

cells.[40] Smoking (nicotine) affects both types of IBD, and it was recently found that these effects are mediated through alpha7 nAChR,[40] a target of *LYNX1*.[39] Furthermore, agonists of alpha7 nAChR affect the disease condition in experimental colitis (a model of IBD) in mice.[41] This evidence suggests a likely role for *LYNX1* in Crohn disease through its effect on nAChR-mediated signaling pathway.

Next, we performed a replication experiment with an independent GWAS data set for Crohn disease[19] and the same eQTL data set. All genes, except *IK*, are replicated at $p < 0.05$ (Table 1). The extent of replication is highly significant ($p = 1.8 \times 10^{-11}$, binomial test).

In summary, of the ten genes predicted by our method at FDR < 0.1, six are known to be associated with Crohn disease and/or IBD or are strong candidates supported by both independent discoveries from literature and replication (*UBE2L3, ORMDL3, PTGER4, LYNX1, EFS*, and *SLC22A5*). One concern is that the genes in the lymphoblasts eQTL data are enriched with immune-related functions,[11] and thus even a randomly chosen set will hit some plausible candidate genes. We performed some estimation and a computational experiment to rule out this possibility. In this eQTL data set, a total of ~9,200 genes have at least one putative eSNP (defined by the same thresholds used for the predicted genes). Assuming there are 1,000 risk genes of Crohn disease (probably an overestimate), and even if they all belong to this list of 9,200 genes, the chance that a randomly chosen gene is associated with Crohn disease is only 1,000/9,200 = 11%, a ratio far below that of our predictions (6 out of 10, $p = 2.5 \times 10^{-4}$, binomial test). Next we randomly sampled two dozen genes from the set of ~9,200 genes and manually judged their relevance to Crohn disease by searching literature in the same way as we did for our candidate genes (Table S2). Only three genes were found to have some immune function, but none specifically related to IBD, compared with seven immunity-related genes (the six candidates plus *IK*) out of ten in our predictions ($p = 0.002$, Fisher's exact test). Despite some limitations (e.g., manual judgment is inherently unreliable), the very conservative nature of our estimation and the strong statistical trends in both analyses suggest that the eQTL data alone cannot explain the results.

Finally, we explore gene networks underlying Crohn disease by using information from *trans* eSNPs that might reveal potential upstream regulatory genes. We found an interesting group of genes with moderate statistical support (nominal p values from .001 to .003; Table S3) that are associated with the same two independent SNPs in *trans* (Figure S2). Most of these genes are related to some aspects of autoimmunity (Table S3). Because they are all affected by the same two eSNPs, this group may be involved in the same molecular pathway affecting Crohn disease. Interestingly, one of the two SNPs, rs10248053, is located inside *RELN* (MIM 600514), which is associated with the age of onset of multiple sclerosis (MIM 126200), an autoimmune disease,[42] and may also be involved in a subtype of T cell leukemia.[43]

Application to Type 2 Diabetes

We next performed an experiment on the GWAS of T2D from DIAGRAM (4,549 cases and 5,579 controls)[44] by using liver eQTL.[45] Previous studies that used the liver data have provided insights on T2D, though only *cis* eSNPs were utilized.[20,45] The p values of the vast majority of the genes closely track the uniform distribution, with about a dozen genes significantly above the diagonal line in the Q-Q plot (Figure 2B). The analysis using permuted GWAS data confirmed the absence of p value inflation (Figure S1). At $p < 10^{-4}$ (FDR 0.3 at this p value

Table 2. Supporting SNPs for Some of the Predicted Genes of Crohn Disease

Gene	SNP	SNP Location	Proximity	eQTL p Value	GWAS p Value	LBF
UBE2L3 (7)	rs2283790	chr22: 20,286,653	*cis*	6.0×10^{-6}	8.9×10^{-6}	6.7
	rs7735799	chr5: 84,248,096	*trans*	3.0×10^{-6}	7.2×10^{-4}	1.3
EFS (16)	rs2154490	chr21: 29,837,833	*trans*	6.0×10^{-6}	3.5×10^{-4}	2.6
	rs8044972	chr16: 75,957,583	*trans*	1.0×10^{-6}	2.0×10^{-2}	1.9
	rs6843282	chr4: 32,433,817	*trans*	9.0×10^{-6}	2.0×10^{-3}	1.4
	rs2210054	chr14: 33,091,890	*trans*	1.0×10^{-5}	5.9×10^{-3}	0.8
LYNX1 (5)	rs921719	chr8: 126,615,379	*trans*	1.0×10^{-5}	7.8×10^{-5}	3.7
	rs11205709	chr1: 50,581,881	*trans*	1.0×10^{-5}	2.8×10^{-3}	1.1
	rs1998564	chr13: 107,034,232	*trans*	1.0×10^{-6}	1.2×10^{-2}	0.7

A SNP is called in *cis* of a gene if it is located within 1 Mb of the transcription start site of this gene. Each SNP in the table actually represents a block of adjacent eSNPs (chosen to be the one with the highest LBF). The number in parentheses in the Gene column shows the total number of putative eSNP blocks (defined as $p < 10^{-5}$ in the eQTL data) of that gene. The LBF column shows the LBF of individual SNP (see Material and Methods). Only SNPs with LBF greater than 0.5 are shown.

cutoff), we predicted four genes (Tables 3 and 4). Two of the four, *PURB* (MIM 608887) and *GNB5* (MIM 604447), are supported only by *trans*-acting SNPs. Similar to Crohn disease, most of the supporting SNPs have only moderate statistical significance in both the GWAS and eQTL data sets.

We found literature evidence supporting three of our predicted genes. *TSPAN8* (MIM 600769) is reported in the original DIAGRAM study and replicated by subsequent GWASs.[44] In an experimental mouse study, haploinsufficiency (deletion of a single copy) of *Gnb5* (the mouse homolog of *GNB5*) caused late-onset obesity, insulin resistance, and liver steatosis on a normal diet, phenotypes strongly resembling human metabolic syndrome and T2D.[46] In addition, a closely related gene, *GNB3* (MIM 139130), is shown to be associated with obesity, insulin resistance, and glucose tolerance in several studies.[47] A SNP close to *JAZF1* (MIM 606246) was associated with T2D,[44,48] and deletion of *Jazf1* in mice leads to increased fat mass and insulin resistance.[49] Interestingly, the SNP rs849134 (close to the supporting *cis* eSNP rs1635852 of *JAZF1*, Table 4) was reported to be strongly associated with both T2D and the expression of *JAZF1* in adipose tissue in the DIAGRAM study.[48] The replication of this finding in liver (rs1635852, associated with T2D and liver expression of *JAZF1*) along with the additional *trans* eSNP from our analysis provides independent evidence of *JAZF1*.

We assembled genomic data from a number of mouse studies to assess the putative role of *PURB* in T2D. The expression of *Purb* in multiple tissues (liver, adipose, muscle, and islet) has been previously found to influence a number of metabolic phenotypes including fat mass, fat to body weight ratio, glucose level, insulin level, glucose to insulin ratio, and oral glucose tolerance test in six mouse crosses via a genetic causality test (FDR 10%).[12,24,25,50] The fact that this relationship was consistently discovered in different tissues and in different genetic backgrounds provides good evidence of a putative role of *PURB* in T2D. We also surveyed genes coexpressed with *PURB* by using previously described tissue-specific coexpression networks constructed from expression data in human and various mouse crosses (see Material and Methods). Genes within a given network module have been found to share similar biological functions. The top gene coexpressed with *PURB* is *SERPINF1* (MIM 172860, appeared in >30% of the 128 coexpression modules containing *PURB*), a strong candidate gene for T2D.[51,52] The strong coexpression between the two genes thus lends further support to *PURB* as a possible T2D-associated gene.

We also analyzed the supporting *trans* eSNPs of the predicted genes of T2D. One supporting eSNP of *PURB*, rs319598 (Table 4), is located in the promoter of *PCBD2* (MIM 609836), which is a cofactor of *HNF1A* (MIM 142410), a liver-specific transcription factor. Mutations of *HNF1A* are responsible for 30%–70% of the cases of maturity-onset diabetes of the young (MODY [MIM 606391]), a rare form of T2D.[53] In another example, we focused on *NDRG2* (MIM 605272), a gene slightly below our threshold ($p = 3.0 \times 10^{-4}$, ranked sixth, not shown in Table 3) but biologically interesting because of its role in β cell protection.[54] One of the two supporting eSNPs of *NDRG2*, rs7334, is close to the 3′ end of *EGFR* (MIM 131550). The *EGFR* signaling pathway plays a key role in pancreatic beta cell development: even a modest attenuation leads to a severe defect in β cells, causing diabetes.[55] These analyses thus further support *PURB* and *NDRG2* as T2D candidate genes and suggest the possible mechanism of their actions, by linking them to the well-established diabetes pathways.

Discussion

We have proposed a general strategy for genetic mapping that integrates information from GWAS with eQTL data.

Table 3. **Top Predicted Genes for Type 2 Diabetes**

Gene	LBF	p Value	GWAS Hit	Supporting Evidence
TSPAN8	6.08	9.4×10^{-6}	yes	associated with T2D in GWAS[48]
PURB	6.07	9.4×10^{-6}	no	expression is causal to T2D-related phenotypes (see text)
GNB5	4.97	8.9×10^{-5}	no	deletion leads to T2D symptoms[46,47]
JAZF1	4.93	9.5×10^{-5}	yes	deletion leads to T2D symptoms;[49] GWAS association[48]

The LBF column is the logarithm of the Bayes Factor for the genes. The p values refer to the p values of LBFs, calculated from simulations. The "GWAS Hit" column shows whether a gene has been implicated as a candidate gene in previous GWAS. See the text for details of the supporting evidence.

Instead of testing individual variants close to a gene, we pool the information in all variants, mostly in *trans*, that perturb the expression of the gene to infer its role in a complex disease. The major difficulties for utilizing *trans* variants to date are their relatively small effects and the pleiotropicity of their influences. We overcome these by utilizing the statistical pattern of multiple *trans* variants, much like a detective using a fingerprint to identify a suspect. Our experiments in Crohn disease and T2D demonstrate the benefits of this approach. Our predictions are often supported by moderate SNPs acting in *trans* (Tables 2 and 4) and thus are not possible to identify by the traditional methods. We believe this general framework significantly extends existing approaches to genetic mapping of complex phenotypes. Both theoretical considerations (complexity of gene regulatory networks) and empirical studies support the importance of *trans* variants affecting gene expression traits. By accessing this large amount of so-far unutilized information, our method could greatly expand our ability to derive new insights from association studies.

Application of our method to two complex diseases led to findings that are supported by multiple lines of evidence. First, for both Crohn disease and T2D, there are clear enrichments of candidate genes from the overall p value distributions (Figure 2). Second, in the study of Crohn disease, we show that all but one of ten predicted genes are replicated in an independent GWAS data set. Third, among our predictions not implicated in human studies before, genetic manipulations of *Efs* and *Gnb5* in experimental mice lead to phenotypes highly consistent with the roles in Crohn disease and T2D, respectively. The collective evidence thus supports our approach as a promising strategy to extract insights from GWASs.

From a population genetic perspective, our strategy provides a significant addition to existing paradigms for genetic mapping. Selection pressure generally keeps proximal, large-effect variants at very low frequencies in the population.[56] This largely explains the observation that most common variants identified in GWASs have small effect sizes.[57] Major efforts are underway to identify rare variants with larger effects by using exome or whole-genome sequencing. Although this has led to a number of discoveries for Mendelian diseases, its success in complex diseases is modest,[58] presumably reflecting large sample size requirements for rare variants. Our strategy is based on the notion that distal variations that only weakly perturb the expression of a gene may survive at significant frequencies in the population because of weaker selection and that there could be many such variations across the genome. Although such weak genetic perturbations may manifest as only modest associations in both the eQTL analysis and the GWAS of a disease, the genes they perturb can play an important role in the disease. In our analysis, although the supporting SNPs of *EFS* and *GNB5* are all in *trans* with modest associations, genetically manipulated mice exhibit severe abnormalities resembling Crohn disease and T2D, respectively.[37,46] Thus our strategy of genetic mapping is capable of finding hidden gene-disease associations by leveraging the collective signals of multiple modest perturbations across the genome.

Our work laid down a general framework for using data from association studies of quantitative molecular traits to interpret GWASs of complex diseases. Many eQTL studies in human have been performed,[9,30] but the vast amount of information in the *trans* eQTL remains unutilized. There are also efforts on mapping QTL of other molecular traits, such as metabolites[59] and epigenetic modifications.[60] Misregulation in these aspects could be important drivers of complex diseases.[61] Our method is readily generalizable to these data sets, by defining the genetic signatures of molecular traits and by matching these signatures to that of the disease.

Our method employs a rigorous statistical framework, leading to major advantages over simple, heuristic methods (e.g., count the number of SNPs shared by the expression and disease traits, defined via some p value threshold). First, both strong and weak loci are taken into account with proper weighting. We incorporate all GWAS associations into the analysis without relying on arbitrary significance thresholds. This is crucial because most of the supporting SNPs in our findings are moderately associated with the expression trait and phenotype but fail to reach genome-wide significance threshold (Tables 2 and 4). Second, our method in general does not permit individual SNPs to dominate the results (see "Model Analysis" in Material and Methods), a common issue when testing association of a set of SNPs in GWASs.[62] Third, the inherent asymmetry of the relationship between the gene and the disease is reflected in our model. A SNP associated with gene expression but not the disease is used as evidence against the gene, and a SNP associated with the disease but not gene expression is treated as noninformative.

An important decision to make when using our tool is the selection of phenotype-appropriate eQTL data sets. For some common diseases, it may be straightforward to

Table 4. Supporting SNPs for the Predicted Genes for Type 2 Diabetes

Gene	SNP	SNP Location	Proximity	eQTL p Value	GWAS p Value	LBF
TSPAN8 (5)	rs7298255	chr12: 69,714,336	*cis*	2.7×10^{-9}	1.3×10^{-6}	6.4
PURB (14)	rs319598	chr5: 134,268,134	*trans*	4.5×10^{-6}	3.9×10^{-5}	4.9
	rs11022347	chr11: 12,391,558	*trans*	7.3×10^{-6}	2.2×10^{-3}	1.3
	rs2028967	chr2: 142,990,138	*trans*	6.4×10^{-6}	5.0×10^{-3}	0.9
GNB5 (7)	rs2021910	chr6: 85,141,735	*trans*	1.8×10^{-6}	3.0×10^{-4}	3.8
	rs13105547	chr4: 52,974,359	*trans*	5.5×10^{-7}	2.1×10^{-2}	1.0
JAZF1 (10)	rs1635852	chr7: 28,155,936	*cis*	3.6×10^{-6}	2.2×10^{-5}	4.8
	rs885720	chr12: 12,139,366	*trans*	2.1×10^{-6}	2.7×10^{-2}	0.6

A SNP is called in *cis* of a gene if it is located within 1 Mb of the transcription start site of this gene. Each SNP in the table actually represents a block of adjacent eSNPs (chosen to be the one with the highest LBF). The number in parentheses in the Gene column shows the total number of putative eSNP blocks (defined as $p < 10^{-5}$ in the eQTL data) of that gene. The LBF column shows the LBF of individual SNP (see Material and Methods). Only SNPs with LBF greater than 0.5 are shown.

choose relevant disease tissues, e.g., brain for psychiatric diseases. The results of our method also provide a quantitative measure of success (the number of significant genes above a certain FDR) and hence an indication of how informative a particular eQTL data set is. In general, we believe that there is probably not a single "correct" tissue; rather, multiple tissues may be informative to different degrees. For instance, it was reported recently that skin, adipose, and blood cells share a substantial fraction of eQTL.[63] Thus, it is possible that skin or blood eQTL may provide information for metabolic diseases. An important future challenge is to characterize the similarity and difference of eQTL across tissues and to develop an analytic framework to integrate information from multiple tissues.

Schadt et al. pioneered the use of eQTL for understanding genetics of complex traits.[12,61] In experimental animals, it is possible to perform genotyping, expression profiling, and phenotyping on the same individuals, allowing researchers to analyze patterns of correlation and dependency among variables to infer causal relationships.[12] It is difficult, however, to apply this strategy to human studies. Expression profiling of disease-related tissues in human subjects is generally costly and often not feasible. Another difficulty is the challenge of mapping *trans* eQTL in human;[9] as a result, most of the existing studies utilizing human eQTL data focus on associations in *cis*,[9,11] with only a few exceptions.[64,65] With our approach, we have circumvented these problems and demonstrated that, even without all three types of data in the same samples, it is possible to infer associations between expression traits and diseases.

In our experiments with T2D, unlike Crohn disease, we predicted a small number of genes at a relatively high error rate (four genes at FDR < 0.3). Several factors might contribute to this lower level of statistical evidence: (1) our T2D analysis is based on liver eQTL data, although more relevant tissues for T2D are probably adipose and pancreatic tissues; and (2) we were unable to perform imputation because of lack of genotype data, thus reducing the overlap between eQTL and GWAS data sets and the power of our method. We believe the results still provide valuable information from existing data. The predictions create a short list of candidate genes; when combining with additional evidence, some interesting hypotheses may emerge. One of our predictions, *PURB*, for instance, is supported by multiple human and mouse genomic data sets, and therefore is an interesting candidate for follow-up analysis.

Although current GWASs typically use only common SNPs measured by array-based genotyping platforms, next-generation sequencing (NGS) technologies provide a window into other types of genetic variation, including rare SNPs, copy-number variants, and indels. Our strategy of matching genetic signatures is based on very general principles and therefore can be easily adapted to these new forms of genetic variations.

Appendix A: The Null and Alternative Distributions of the Association Statistics of SNPs

For a given SNP, let T be its association test statistic (p value can be converted to T, assuming a standard normal distribution) for a trait, either an expression trait or phenotype. We use M to denote the model: $M = 0$ corresponds to the null model of no association between the SNP and the trait, and $M = 1$ corresponds to the alternative model. We need to compute the probability distribution, $P(T \mid M)$. In the null model, typically, $P(T \mid M = 0)$ follows the standard normal distribution. In the alternative model, the distribution $P(T \mid M = 1)$ depends on the statistical test through which T is derived and the effect size of the SNP, as explained below.

Binary Trait

We assume the Armitage trend test was used to derive the test statistics. We follow the notations in Slager and Schaid:[66] for a given SNP, there are two alleles, A for the high-risk allele and a for the other. The frequency of the

i^{th} genotype ($i = 0,1,2$ is the number of A alleles) in cases is p_i, and its frequency in controls is q_i. The number of case and control individuals with each genotype are shown in Table A1.

Table A1. The Number of Individuals with Given Genotypes

Genotype	aa	aA	AA	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

We define a variable $x_i = i$, the number of A alleles in a genotype, the Armitage trend test can be written as $T = U/\sqrt{Var(U)}$, where

$$U = \sum_i x_i \left(\frac{S}{N} r_i - \frac{R}{N} s_i \right). \qquad \text{(Equation A1)}$$

Under the null hypothesis of no association, the mean of $U(\mu_0)$ is 0 and its variance is given by[66]

$$\sigma_0^2 = N\phi(1-\phi) \left[\sum_i x_i^2 q_i - \left(\sum_i x_i q_i \right)^2 \right], \qquad \text{(Equation A2)}$$

where $\phi = R/N$ is the fraction of cases in the sample. Based on the central limit theorem, the asymptotic distribution, $T = U/\sigma_0 \sim N(0,1)$.

Under the alternative hypothesis of association, the mean and variance of U are given by[66]

$$\mu_1 = N\phi(1-\phi) \sum_i x_i (p_i - q_i) \qquad \text{(Equation A3)}$$

$$\sigma_1^2 = N\phi(1-\phi)^2 \left[\sum_i x_i^2 p_i - \left(\sum_i x_i p_i \right)^2 \right] + N\phi^2(1-\phi) \times \left[\sum_i x_i^2 q_i - \left(\sum_i x_i q_i \right)^2 \right]. \qquad \text{(Equation A4)}$$

Thus, T follows normal distribution:

$$T = \frac{U}{\sigma_0} \sim N\left(\frac{\mu_1}{\sigma_0}, \frac{\sigma_1^2}{\sigma_0^2} \right). \qquad \text{(Equation A5)}$$

The parameters p_i and q_i are unknown, but they are related to the effect size and allele frequency of the SNP in the population. For the i^{th} genotype, let f_i be its penetrance (the probability of disease given the genotype) and g_i be its frequency in the population. Then p_i and q_i are related to f_i and g_i through the Bayes theorem:

$$p_i = \frac{f_i g_i}{\sum_i f_i g_i}, \quad q_i = \frac{(1-f_i) g_i}{\sum_i (1-f_i) g_i}. \qquad \text{(Equation A6)}$$

The sum $\Sigma_i f_i g_i$ is also called disease prevalence, K. Assuming Hardy-Weinberg equilibrium, the genotype frequencies are simply related to p, the frequency of the risk allele (assumed known, e.g., from HapMap). Suppose the effect size (the logarithm of the odds ratio of the risk allele) is β, the relative risk is then approximately $\gamma = e^{\beta}$. Assuming the multiplicative genetic model, we have $f_1 = \gamma f_0$ and $f_2 = \gamma^2 f_0$. With the fact $K = \Sigma_i f_i g_i$, we could solve f_0:

$$f_0 = \frac{K}{[\gamma p + (1-p)]^2} \qquad \text{(Equation A7)}$$

In summary, given K and p, p_i and q_i can be expressed as functions of β. Plug in p_i and q_i to equations of μ_1, σ_1, we can write μ_1, σ_1 as functions of β as well: $\mu_1(\beta; K, p)$ and $\sigma_1(\beta; K, p)$. Because β is unknown, we assume a prior distribution of β: $\beta | M = 1 \sim N(0, \sigma_a^2)$ and integrate out β in the distribution $P(T \mid M = 1)$:

$$\begin{aligned} P(t \mid M = 1) &= \int P(t \mid \beta) P(\beta \mid M = 1) d\beta \\ &= \int N\left(t \mid \frac{\mu_1(\beta)}{\sigma_0}, \frac{\sigma_1^2(\beta)}{\sigma_0} \right) N(\beta \mid 0, \sigma_a^2) d\beta, \end{aligned} \qquad \text{(Equation A8)}$$

Where $N(t|.)$ is the p.d.f. of the normal distribution. A common choice of the prior distribution is $\sigma_a^2 = 0.2$.[17]

Quantitative Trait

For an association study of a quantitative trait, we assume a z-test of the linear regression coefficient is used. Specifically, we have the following regression:

$$y_i = \mu + \beta x_i + \varepsilon_i, \qquad \text{(Equation A9)}$$

where x_i and y_i are the genotype and phenotype of the i^{th} subject in the sample, respectively, and $\varepsilon_i \sim N(0,\sigma^2)$ is the error term. The statistic test of whether $\beta = 0$ is: $T = b_1/\sigma(b_1)$, where b_1 is the MLE of β. The variance of b_1 is given by

$$\sigma^2(b_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{2Np(1-p)}, \qquad \text{(Equation A10)}$$

where N is sample size and p is the allele frequency (assuming HWE). T follows standard normal distribution under $M = 0$: $T = b_1/\sigma(b_1) \sim N(0,1)$. In this calculation, we assume that σ^2 is known. If this is not true, we need to replace σ^2 with its MLE, σ^2 in T. However, with large sample size (which is often the case in large GWASs), the two are close, so for simplicity, we will use σ^2 instead. Under $M = 1$, we assume the prior distribution $\beta \sim N(0, \sigma_a^2 \sigma^2)$, according to Servin and Stephens,[14] where σ_a reflects the typical effect size compared with the standard deviation of the quantitative trait. With the distribution $b_1 \sim N(\beta, \sigma^2(b_1))$, we have

$$\begin{aligned} P(b_1 \mid M = 1) &= \int P(b_1 \mid \beta) P(\beta \mid M = 1) d\beta \\ &= \int N(b_1 \mid \beta, \sigma^2(b_1)) N(\beta \mid 0, \sigma_a^2 \sigma^2) d\beta \end{aligned} \qquad \text{(Equation A11)}$$

It is easy to show that b_1 under $M = 1$ follows the normal distribution $b_1 \sim N(0, \sigma^2(b_1) + \sigma_a^2\sigma^2)$, thus

$$T = \frac{b_1}{\sigma(b_1)} \sim N\left(0, 1 + 2Np(1-p)\sigma_a^2\right). \qquad \text{(Equation A12)}$$

The hyperparameter of the prior distribution, σ_a, is fixed at 0.5 in our experiments, based on the earlier study of Bayesian association mapping of quantitative traits.[14]

Appendix B: Relating the Bayes Factors of Genes to the SNP-Level Bayes Factors in eQTL and GWAS Data

Under the null hypothesis of no gene-disease relationship (Z = 0), the eQTL and GWAS data are independent (Figure 1D), and we thus have the model evidence of the null hypothesis:

$$P(x, y \mid H_0) = \prod_{i=1}^{N} P(x_i \mid H_0) P(y_i \mid H_0) = \prod_{i=1}^{N} f_0(x_i) g_0(y_i), \qquad \text{(Equation A13)}$$

where the two functions describe the probability of eQTL and GWAS summary statistics, respectively:

$$f_0(x_i) = P(x_i \mid H_0) = (1-\alpha)P(x_i \mid U_i = 0) + \alpha P(x_i \mid U_i = 1) \qquad \text{(Equation A14)}$$

$$g_0(y_i) = P(y_i \mid H_0) = (1-\beta)P(y_i \mid V_i = 0) + \beta P(y_i \mid V_i = 1). \qquad \text{(Equation A15)}$$

Under the alternative hypothesis (Z = 1), we plug in the relevant terms to Equation 3:

$$P(x_i, y_i \mid H_1) = (1-\alpha)P(x_i \mid U_i = 0) g_0(y_i) + \alpha P(x_i \mid U_i = 1) P(y_i \mid V_i = 1). \qquad \text{(Equation A16)}$$

Divide the model evidence of H_1 at the i^{th} SNP over that of H_0, we have

$$B_i = \frac{P(x_i, y_i \mid H_1)}{P(x_i, y_i \mid H_0)} = \frac{(1-\alpha)P(x_i \mid U_i = 0) g_0(y_i) + \alpha P(x_i \mid U_i = 1) P(y_i \mid V_i = 1)}{f_0(x_i) g_0(y_i)}. \qquad \text{(Equation A17)}$$

Eliminate the common terms in the numerator and the denominator, and we have Equation 5 shown in the main text. The gene level BF is simply the product of B_is.

Supplemental Data

Supplemental Data include two figures and three tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

The authors wish to thank Kathleen Giacomini, Sook Wah Yee, Saunak Sen, Mark Seielstad, Neil Risch, and Jiashun Zheng for helpful discussions and technical support. We would also like to thank the DIAGRAM consortium for providing the T2D GWAS data and Mark McCarthy for valuable comments on the manuscript. This work was supported by National Institutes of Health grants R01(GM070808) and U19(GM61390), the NIH Center for Systems and Synthetic Biology (P50 GM081879), and a Packard Fellowship in Science and Engineering to H.L.

Received: October 10, 2012
Revised: March 7, 2013
Accepted: March 25, 2013
Published: May 2, 2013

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes, http://browser.1000genomes.org
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org/
Sherlock, http://sherlock.ucsf.edu/

References

1. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.
2. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al.; ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
3. Vidal, M., Cusick, M.E., and Barabási, A.L. (2011). Interactome networks and human disease. Cell *144*, 986–998.
4. Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.
5. Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2008). Regulatory changes underlying expression differences within and between *Drosophila* species. Nat. Genet. *40*, 346–350.
6. Brem, R.B., and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci. USA *102*, 1572–1577.
7. Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R., and Kruglyak, L. (2003). Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nat. Genet. *35*, 57–64.
8. Cheung, V.G., Nayak, R.R., Wang, I.X., Elwyn, S., Cousins, S.M., Morley, M., and Spielman, R.S. (2010). Polymorphic cis- and trans-regulation of human gene expression. PLoS Biol. *8*, e1000480.
9. Montgomery, S.B., and Dermitzakis, E.T. (2011). From expression QTLs to personalized transcriptomics. Nat. Rev. Genet. *12*, 277–282.
10. Price, A.L., Patterson, N., Hancks, D.C., Myers, S., Reich, D., Cheung, V.G., and Spielman, R.S. (2008). Effects of cis and trans genetic ancestry on gene expression in African Americans. PLoS Genet. *4*, e1000294.
11. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to

be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.

12. Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet. *37*, 710–717.
13. Duan, S., Huang, R.S., Zhang, W., Bleibel, W.K., Roe, C.A., Clark, T.A., Chen, T.X., Schweitzer, A.C., Blume, J.E., Cox, N.J., and Dolan, M.E. (2008). Genetic architecture of transcript-level variation in humans. Am. J. Hum. Genet. *82*, 1101–1113.
14. Servin, B., and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet. *3*, e114.
15. Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (Cambridge, UK: Cambridge University Press).
16. Liu, J.Z., McRae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., Hayward, N.K., Montgomery, G.W., Visscher, P.M., Martin, N.G., and Macgregor, S.; AMFS Investigators. (2010). A versatile gene-based test for genome-wide association studies. Am. J. Hum. Genet. *87*, 139–145.
17. Stephens, M., and Balding, D.J. (2009). Bayesian statistical methods for genetic association studies. Nat. Rev. Genet. *10*, 681–690.
18. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.
19. Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. *42*, 1118–1125.
20. Zhong, H., Yang, X., Kaplan, L.M., Molony, C., and Schadt, E.E. (2010). Integrating pathway analysis and genetics of gene expression for genome-wide association studies. Am. J. Hum. Genet. *86*, 581–591.
21. Greenawalt, D.M., Dobrin, R., Chudin, E., Hatoum, I.J., Suver, C., Beaulaurier, J., Zhang, B., Castro, V., Zhu, J., Sieberts, S.K., et al. (2011). A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. Genome Res. *21*, 1008–1016.
22. Yang, X., Zhang, B., Molony, C., Chudin, E., Hao, K., Zhu, J., Gaedigk, A., Suver, C., Zhong, H., Leeder, J.S., et al. (2010). Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. Genome Res. *20*, 1020–1036.
23. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–428.
24. Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., MacNeil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K., et al. (2008). Variations in DNA elucidate molecular networks that cause disease. Nature *452*, 429–435.
25. Derry, J.M., Zhong, H., Molony, C., MacNeil, D., Guhathakurta, D., Zhang, B., Mudgett, J., Small, K., El Fertak, L., Guimond, A., et al. (2010). Identification of genes and networks driving cardiovascular and metabolic phenotypes in a mouse F2 intercross. PLoS ONE *5*, e14319.
26. Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. *4*, e17.
27. Wakefield, J. (2008). Reporting and interpretation in genome-wide association studies. Int. J. Epidemiol. *37*, 641–653.
28. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B Met. *57*, 289–300.
29. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., et al.; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. *40*, 955–962.
30. Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. Nat. Rev. Genet. *10*, 184–194.
31. Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997–1004.
32. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A.; 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
33. Lees, C.W., Barrett, J.C., Parkes, M., and Satsangi, J. (2011). New IBD genetics: common pathways with other diseases. Gut *60*, 1739–1753.
34. Zhernakova, A., Stahl, E.A., Trynka, G., Raychaudhuri, S., Festen, E.A., Franke, L., Westra, H.J., Fehrmann, R.S., Kurreeman, F.A., Thomson, B., et al. (2011). Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet. *7*, e1002004.
35. Cotsapas, C., Voight, B.F., Rossin, E., Lage, K., Neale, B.M., Wallace, C., Abecasis, G.R., Barrett, J.C., Behrens, T., Cho, J., et al.; FOCiS Network of Consortia. (2011). Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet. *7*, e1002254.
36. Donlin, L.T., Roman, C.A., Adlam, M., Regelmann, A.G., and Alexandropoulos, K. (2002). Defective thymocyte maturation by transgenic expression of a truncated form of the T lymphocyte adapter molecule and Fyn substrate, Sin. J. Immunol. *169*, 6900–6909.
37. Donlin, L.T., Danzl, N.M., Wanjalla, C., and Alexandropoulos, K. (2005). Deficiency in expression of the signaling protein Sin/Efs leads to T-lymphocyte activation and mucosal inflammation. Mol. Cell. Biol. *25*, 11035–11046.
38. Muraoka, M., Hasegawa, H., Kohno, M., Inoue, A., Miyazaki, T., Terada, M., Nose, M., and Yasukawa, M. (2006). IK cytokine ameliorates the progression of lupus nephritis in MRL/lpr mice. Arthritis Rheum. *54*, 3591–3600.
39. Moriwaki, Y., Yoshikawa, K., Fukuda, H., Fujii, Y.X., Misawa, H., and Kawashima, K. (2007). Immune system expression of SLURP-1 and SLURP-2, two endogenous nicotinic acetylcholine receptor ligands. Life Sci. *80*, 2365–2368.
40. Galitovskiy, V., Qian, J., Chernyavsky, A.I., Marchenko, S., Gindi, V., Edwards, R.A., and Grando, S.A. (2011). Cytokine-induced alterations of α7 nicotinic receptor in colonic CD4 T cells mediate dichotomous response to nicotine in murine

models of Th1/Th17- versus Th2-mediated colitis. J. Immunol. *187*, 2677–2687.

41. Snoek, S.A., Verstege, M.I., van der Zanden, E.P., Deeks, N., Bulmer, D.C., Skynner, M., Lee, K., Te Velde, A.A., Boeckxstaens, G.E., and de Jonge, W.J. (2010). Selective alpha7 nicotinic acetylcholine receptor agonists worsen disease in experimental colitis. Br. J. Pharmacol. *160*, 322–333.
42. Baranzini, S.E., Wang, J., Gibson, R.A., Galwey, N., Naegelin, Y., Barkhof, F., Radue, E.W., Lindberg, R.L., Uitdehaag, B.M., Johnson, M.R., et al. (2009). Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Hum. Mol. Genet. *18*, 767–778.
43. Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S.L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M., et al. (2012). The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature *481*, 157–163.
44. Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G., et al.; Wellcome Trust Case Control Consortium. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat. Genet. *40*, 638–645.
45. Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008). Mapping the genetic architecture of gene expression in human liver. PLoS Biol. *6*, e107.
46. Wang, Q., Levay, K., Chanturiya, T., Dvoriantchikova, G., Anderson, K.L., Bianco, S.D., Ueta, C.B., Molano, R.D., Pileggi, A., Gurevich, E.V., et al. (2011). Targeted deletion of one or two copies of the G protein β subunit Gβ5 gene has distinct effects on body weight and behavior in mice. FASEB J. *25*, 3949–3957.
47. Kopf, D., Cheng, L.S., Blandau, P., Hsueh, W., Raffel, L.J., Buchanan, T.A., Xiang, A.H., Davis, R.C., Rotter, J.I., and Lehnert, H. (2008). Association of insulin sensitivity and glucose tolerance with the c.825C>T variant of the G protein beta-3 subunit gene. J. Diabetes Complications *22*, 205–209.
48. Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G., et al.; MAGIC investigators; GIANT Consortium. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat. Genet. *42*, 579–589.
49. Langberg, K.A., Ma, L., Sharma, N.K., Hanis, C.L., Elbein, S.C., Hasstedt, S.J., and Das, S.K.; American Diabetes Association GENNID Study Group. (2012). Single nucleotide polymorphisms in JAZF1 and BCL11A gene are nominally associated with type 2 diabetes in African-American families from the GENNID study. J. Hum. Genet. *57*, 57–61.
50. Yang, X., Deignan, J.L., Qi, H., Zhu, J., Qian, S., Zhong, J., Torosyan, G., Majid, S., Falkard, B., Kleinhanz, R.R., et al. (2009). Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. Nat. Genet. *41*, 415–423.
51. Crowe, S., Wu, L.E., Economou, C., Turpin, S.M., Matzaris, M., Hoehn, K.L., Hevener, A.L., James, D.E., Duh, E.J., and Watt, M.J. (2009). Pigment epithelium-derived factor contributes to insulin resistance in obesity. Cell Metab. *10*, 40–47.
52. Böhm, A., Ordelheide, A.M., Machann, J., Heni, M., Ketterer, C., Machicao, F., Schick, F., Stefan, N., Fritsche, A., Häring, H.U., and Staiger, H. (2012). Common genetic variation in the SERPINF1 locus determines overall adiposity, obesity-related insulin resistance, and circulating leptin levels. PLoS ONE *7*, e34035.
53. McKusick, V.A. (1998). Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders (Baltimore: John Hopkins University Press).
54. Shen, L., Liu, X., Hou, W., Yang, G., Wu, Y., Zhang, R., Li, X., Che, H., Lu, Z., Zhang, Y., et al. (2010). NDRG2 is highly expressed in pancreatic beta cells and involved in protection against lipotoxicity. Cell. Mol. Life Sci. *67*, 1371–1381.
55. Miettinen, P.J., Ustinov, J., Ormio, P., Gao, R., Palgi, J., Hakonen, E., Juntti-Berggren, L., Berggren, P.O., and Otonkoski, T. (2006). Downregulation of EGF receptor signaling in pancreatic islets causes diabetes due to impaired postnatal beta-cell growth. Diabetes *55*, 3299–3308.
56. Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. *69*, 124–137.
57. Ku, C.S., Loy, E.Y., Pawitan, Y., and Chia, K.S. (2010). The pursuit of genome-wide association studies: where are we now? J. Hum. Genet. *55*, 195–206.
58. Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al. (2012). Exome sequencing and the genetic basis of complex traits. Nat. Genet. *44*, 623–630.
59. Suhre, K., Shin, S.Y., Petersen, A.K., Mohney, R.P., Meredith, D., Wägele, B., Altmaier, E., Deloukas, P., Erdmann, J., Grundberg, E., et al.; CARDIoGRAM. (2011). Human metabolic individuality in biomedical and pharmaceutical research. Nature *477*, 54–60.
60. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J., et al. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. *6*, e1000952.
61. Schadt, E.E. (2009). Molecular networks as sensors and drivers of common human diseases. Nature *461*, 218–223.
62. Wang, K., Zhang, H., Kugathasan, S., Annese, V., Bradfield, J.P., Russell, R.K., Sleiman, P.M., Imielinski, M., Glessner, J., Hou, C., et al. (2009). Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn disease. Am. J. Hum. Genet. *84*, 399–405.
63. Grundberg, E., Small, K.S., Hedman, A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.P., Meduri, E., Barrett, A., et al.; Multiple Tissue Human Expression Resource (MuTHER) Consortium. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet. *44*, 1084–1089.
64. Fehrmann, R.S., Jansen, R.C., Veldink, J.H., Westra, H.J., Arends, D., Bonder, M.J., Fu, J., Deelen, P., Groen, H.J., Smolonska, A., et al. (2011). Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. *7*, e1002197.
65. Small, K.S., Hedman, A.K., Grundberg, E., Nica, A.C., Thorleifsson, G., Kong, A., Thorsteindottir, U., Shin, S.Y., Richards, H.B., Soranzo, N., et al.; GIANT Consortium; MAGIC Investigators; DIAGRAM Consortium; MuTHER Consortium. (2011). Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat. Genet. *43*, 561–564.
66. Slager, S.L., and Schaid, D.J. (2001). Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. Hum. Hered. *52*, 149–153.