Nucleic Acids Research Advance Access published March 24, 2012

Nucleic Acids Research, 2012, 1–10
doi:10.1093/nar/gks269

Inference of modules associated to eQTLs

Anat Kreimer[1,*], Oren Litvin[2], Ke Hao[3], Cliona Molony[3], Dana Pe'er[2] and Itsik Pe'er[4]

[1]Department of Biomedical Informatics, Columbia University, New York 10032, [2]Department of Biological Sciences, Columbia University, New York 10027, NY, [3]Merck Research Labs, 33 Ave. Louis Pasteur, Boston, MA 02115, USA and [4]Department of Computer Science, Columbia University, New York 10027, NY

Received November 8, 2011; Revised February 7, 2012; Accepted March 9, 2012

ABSTRACT

Cataloging the association of transcripts to genetic variants in recent years holds the promise for functional dissection of regulatory structure of human transcription. Here, we present a novel approach, which aims at elucidating the joint relationships between transcripts and single-nucleotide polymorphisms (SNPs). This entails detection and analysis of modules of transcripts, each weakly associated to a single genetic variant, together exposing a high-confidence association signal between the module and this 'main' SNP. To explore how transcripts in a module are related to causative loci for that module, we represent such dependencies by a graphical model. We applied our method to the existing data on genetics of gene expression in the liver. The modules are significantly more, larger and denser than found in permuted data. Quantification of the confidence in a module as a likelihood score, allows us to detect transcripts that do not reach genome-wide significance level. Topological analysis of each module identifies novel insights regarding the flow of causality between the main SNP and transcripts. We observe similar annotations of modules from two sources of information: the enrichment of a module in gene subsets and locus annotation of the genetic variants. This and further phenotypic analysis provide a validation for our methodology.

INTRODUCTION

Variation in genomic DNA can affect function in multiple ways, most typically by alteration of the expressed quantity or sequence content of local transcripts. This premise motivated extensive studies over the last decade, cataloging the influence of human genetic variants on gene expression, most often in *cis* (1,2). Local gene expression level is formally considered as a quantitative trait that is directly modified by allelic variation in regulatory elements (3,4). Such modifications of transcriptional regulation have been documented to affect health-related traits as diverse as asthma (5) and low density lipoprotein (LDL) cholesterol concentration (6).

Yet, for large fraction of single-nucleotide polymorphisms (SNPs) with well supported associations to disease phenotypes (7) which are neither coding, nor linked to coding SNPs in *cis*, no *cis*-regulatory effect have been reported in studies conducted thus far. A compelling biological hypothesis is that such a SNP does change the transcriptome state or program in order to exert its phenotypic impact, and this regulation is mediated by a transcript in *cis*, but in the particular tissue examined, the changes to transcription level of the mediator gene are too minute to guarantee detection in small association cohorts. This hypothesis leads to an approach for mapping expression quantitative trait loci (eQTLs) that is focused on downstream effects of a regulatory SNP across multiple genes in *trans*, rather than the *cis*-transcript that may mechanistically mediate the effect. A related approach had been successful in simpler organisms (8), motivating this work.

Data on both gene expression and SNP variation across multiple individuals, often termed genetic genomics have facilitated identification of thousands of expression single-nucleotide polymorphisms (eSNPs) (9,10). Approaches that combine these two types of data along with additional factors including the previously inferred biological network structure (11), modularity of gene expression (12), pathway analysis (13) and enzymatic activity (14) had been proposed. However, tying genetic variation in specific loci to phenotypes is still an active field of research.

In this study, we focus on the modularity of gene regulatory networks, a major organizing principle of biological systems (15). A module is the fundamental unit of a biological network that consists of a set of elements (e.g. genes) working jointly to fulfill a distinct function. Several studies have used this property to gain better understanding of the regulatory mechanisms (16) that are affected by genetic variation. Litvin *et al.* (8) characterize how genetic variants in multiple loci combine to influence the expression of clusters of co-expressed genes

*To whom correspondence should be addressed. Tel: +1 212 939 7135; Fax: +1 212 666 0140; Email: anat.kreimer@gmail.com

in yeast. Ghazalpour *et al.* (12) used co-expression networks to study the genetics of complex physiological traits that are relevant to the metabolic syndrome. Schadt *et al.* (11) used previously reconstructed regulatory networks of genes in mouse and human (17) to support the existing Genome Wide Association Studies (GWAS) results. Known pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) were used by Zhong *et al.* (13) for the same purpose. Common to all these studies are three steps. The first two are independent: (i) construction of a network from gene expression data; and (ii) detection of association between genetic variants and expression traits; the final step is (iii) integration of genetic association into the network.

However, it is artificial to separate the stages of network construction based on expression data only from a single SNP–transcript association mapping. Ideally, one would combine information from multiple transcripts with genetics in a unified analysis. This motivates complementary approaches to analysis of eSNPs. Specifically, our premise is that the modular organization of gene regulation can be used to pinpoint eSNPs that affect multiple, rather than single genes. Therefore, we developed a method that focuses on groups of transcripts (modules) that are each associated with a single genetic variant.

We present a novel approach that entails analyzing modules of transcripts, each associated to a single genetic variant. These modules are constructed based on both available types of data: transcript expression and genotypes. We combine these transcripts into modules that each share an associated SNP, which we denote as the 'main' SNP of that module. This step utilizes the modular organization of gene regulation. We filter the modules according to a confidence score. This score allows us to identify groups of transcripts that are associated to a SNP even if their individual association is not genome-wide significant. We examine the topology of modules, accounting for independent co-association, which is not merely the result of co-expression. This step allows us to infer the flow of causality between the main SNP and the transcripts in the module. We distinguish direct versus indirect SNP–transcript associations through another intermediate transcript whose expression level is co-associated to the same SNP. The main SNP can possibly have *cis*- or *trans*-effects on the transcripts in the module. A local *cis*-effect on a transcript that is either included or excluded from a module can in turn have a modular *trans*-regulatory effect on the other transcripts in the module by virtue of its changed expression levels or altered produced protein (e.g. a mutation in transcription factor).

Regulatory effects can be categorized by *cis*- and *trans*-effects. The *cis*-effects of eSNPs are often due to changes within the promoter, enhancer or other regulatory regions of a gene that may change the expression of that gene. *Trans*-effects of the main SNP on module transcripts can be the outcome of two potentially overlapping scenarios: First, a *cis* main SNP that is located within or close by the coding region of one of the genes in the module can alter the produced protein. The altered protein may then have a *trans*-regulatory effect on the other transcripts in the module by virtue of its differential expression level despite the protein itself being potentially unmodified. Second, a *trans* main SNP that is located within or close by the coding region of a gene that is not a part of the module can alter the produced protein. This distant altered protein may then have a *trans*-effect on the other transcripts in the module by virtue of its modified sequence, despite potentially maintaining its expression level.

All methods previously introduced group transcripts by a shared associated marker and determine intra-cluster interactions by using the correlation of gene expression levels. To our knowledge, this is the first work where a confidence score is assigned to each module and direct/indirect interactions are determined between pairs of transcripts within a module illustrating the dependence/independence of their expression levels conditioned on the main SNP. We are thus able to go beyond traditional clustering-related methods that are based on expression only, and in fact, examine the joint association and the topology of the modules and not merely their content. For completion, we further search for regulatory hierarchical structure within each module: we examine SNPs whose association to transcript levels in a module is conditioned on the main SNP, and denote those as 'secondary' SNPs. This step is illustrated as a decision tree where samples in each module are split, first by the genotype of the main SNP and then by the genotype of the secondary SNP. We applied our method to data regarding genotype and gene expression in the liver across 371 samples. This data had been previously analyzed in other means (11). We observe known relationships from the literature between a module and its associated genetic variants, thereby providing support to our methodology.

MATERIALS AND METHODS

Data details and processing

The DeLiver data set by Merck had been described elsewhere (11). Briefly, the raw data set consists of 653 894 SNPs and 25 917 expression probes (log-transformed values) with an Entrez gene ID assayed for 385 samples. We remove 99 expression probes that are mapped to the Y chromosome. Multiple probes that are mapped to the same gene had been averaged if correlated ($r > 0.75$) or discarded otherwise, resulting in 18 883 genes with unique Entrez IDs. 5055 genes had variable levels of liver expression across the individuals (SD > 0.2). Standard filters have been applied to the SNP data: Minor allele frequency >0.05, SNP missingness rate <0.1 and individual missingness rate <0.1 (18). After filtering, the data for analysis consists of 371 samples (200 males, 171 females) with 557 456 SNPs and 5055 genes.

For each individual i, we denote the expression levels of each transcript t by $X(i,t)$, and the genotype for each SNP s by $G(i,s)$.

Step 1—nominal association testing

We test for association between pairs (s,t) of any SNP s and transcript t using linear regression and record the results between every (s,t) pair with nominal

$P < 10^{-5}$. To eliminate transcripts whose association statistic is strongly distorted, we repeated the analysis 1000 times with permuted data, obtained by randomly switching the samples' labels, discarding recurrently observed transcripts as follows. A small fraction of observed association pairs tend to recur in permuted data sets more than expected (Supplementary Table S1). Specifically, 2979 of the observed association pairs detected in the real data appear exactly once in the 1000 permuted data sets (<676 expected), and 520 recur twice (<7 expected). This suggests a bias in the test statistic for these pairs, and we discard all 623 pairs that appear in two permutations or more from subsequent analysis.

When considering association pairs detected in the real and permuted data, we note that over dispersion of the test-statistic exists in both. In the real data, $10^{-4.61}$ of (s,t) pairs attain a test statistic theoretically corresponding to a $P = 10^{-5}$ (Supplementary Figure S1), $\lambda = 1.017$, whereas in the 100 permutations using all SNPs in the data, only $10^{-4.65}$ of such pairs attain this level, $\lambda = 1.0485$. We use the nominal $P = 10^{-5}$ as a threshold, keeping in mind that this P-value is not genome-wide significant, and 69 172 random association pairs are expected to pass this threshold by chance alone. This justifies the use of such a threshold, as our methodology relies on having a variety of association pairs, that only when cross-compared across transcripts would yield a meaningful result.

Step 2—module construction, scoring and filtering

The putatively associated transcripts are binned by their SNP s, each bin hereby referred to as a *module*. This associated SNP s is denoted as the 'main' SNP. We consider each module in turn. Let M be a module of size k, with a set of transcripts $\{t_1,\ldots,t_k\}$ and a main SNP s. For each transcript t_i we consider the P-value denoted $Pval(t_i)$ of the association test between the main SNP s and its expression level. We compute the empirical false positive rate (EFPR) for each such P-value by permutation: We use 100 permutations to tally the average number of P-values better than $Pval(t_i)$ across the permuted data sets divided by the analogous number in the real data. This ratio is the EFPR corresponding to $Pval(t_i)$. We follow a similar procedure to calculate the analogous ratio for module size k: EFPR(k) is defined as the ratio of the average number of modules with size bigger than k across the permuted data sets and the analogous number in the real data. The score $S(M)$ of the module M

$$S(M) = -\sum_{i=1}^{k} \log(\mathrm{EFPR}(p\mathrm{val}(t_i))) - \log(\mathrm{EFPR}(k))$$

is justified as a log-likelihood-ratio that compares two hypotheses (Supplementary Text S1). In order to assign significance to the obtained scores, we again use 100 permutations. We score each of the modules in the permuted data sets against the other 99 (a 'leave one out' procedure) in a similar process to the one described for computing the scores of modules in the real data. We thereby provide an empirical P-value interpretation by scaling the scores of modules in the real data, compared with the average score of modules in permutations, i.e. the true positive rate (TPR) of the score of a module.

Step 3—finding secondary SNPs

We split the samples by the genotype of the main SNP into three subsets of samples with genotypes AA, Aa and aa, respectively (where A and a are the major and minor alleles, respectively). AA and Aa are the two larger subsets of samples. In each of those two subsets, we then turn to find the corresponding two subset-specific SNPs that best explain the expression of the largest group of genes in each subset, and denote these 'secondary' SNPs (8,19). To search for secondary SNPs, we test each SNP for association only to the transcripts within the module, and only within the current subset of samples. We discard pairs of transcript and SNP in recurrently observed association pairs by using 1000 permutations and removing all association pairs that appear in one permutation or more (empirical FDR < 0.001). We consider all SNPs that comply with three criteria: (i) maximal-size subgroup of transcripts (with minimum of five transcripts), (ii) F-test for independent association of transcript pairs and (iii) minimal product of association P-values. More specifically: For each module, and each genotype group we first list all SNPs that achieve an association nominal P-value of 10^{-5} or better with a large subgroup of transcripts (five transcripts or more). We consider only those whose subgroup is maximal as candidate secondary SNPs. We test all possible pairs of transcripts in the subgroup for conditional association (See 'Analysis of dependencies within modules' section), and discard a candidate secondary SNP if any pair fails the test. Out of this list, we seek the SNP with the minimal product of association P-values with its subgroup of transcripts. These steps control for false discovery, because the phenomena of big and edge dense modules does not exist in permutations.

Analysis of dependencies within modules

For each module, we consider all possible ordered triplets $(t,t's)$ of two transcripts t, t' whose levels are significantly associated with the same main SNP s. We define bi-directional triplets where association is mutually independent, i.e. for both association pairs remain nominally significant given the respective other transcript versus 'uni-directional' triplets where association is directionally independent (Figure 1a). Formally, we test whether the association model provides significantly better fit to the data than the null model.

$$\begin{aligned} &\text{Null model:} && X(i,t) = \alpha_0 + \alpha_1 \cdot X(i,t') + \varepsilon_1 \\ &\text{Association model:} && X(i,t) = \beta_0 + \beta_1 \cdot X(i,t') + \beta_2 \cdot G(i,s) + \varepsilon_2 \end{aligned}$$

We use the F-test for better fit symmetrically, attempting to explain the expression levels of either t by t' or the converse, with or without genotypes (testing the significance of β_2 being non-zero coefficient would yield the same results). We describe independence of associations in each module M as a graph $G(M)$, whose vertices correspond to transcripts. A directed/bidirectional edge

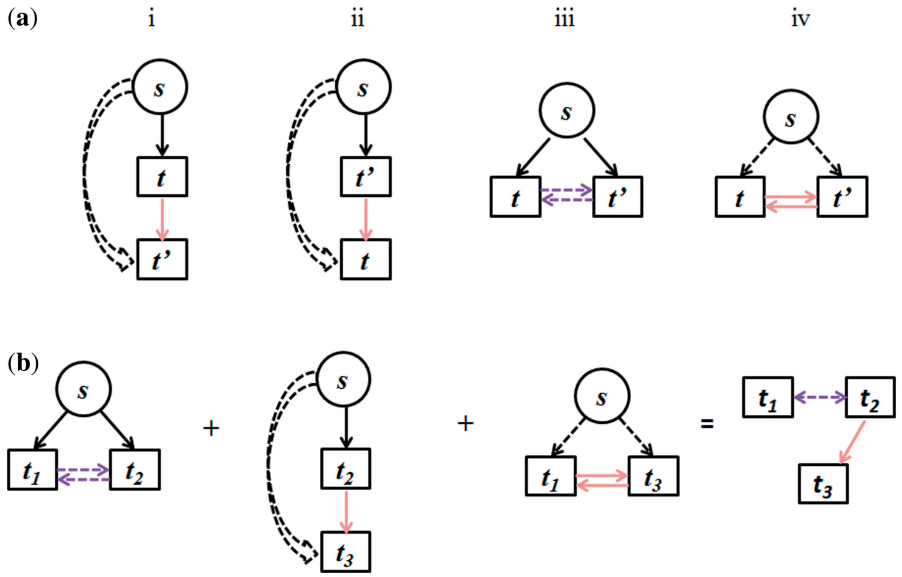

Figure 1. (**a**) Graphical illustration of a triplet with two transcripts t and t' and a main SNP s. The dashed/full black line represents dependent/independent association between a SNP and a transcript, respectively. The uni/bi-directional pink/purple line represents an edge that connects transcripts with directionally/mutually independent association to the main SNP (i) unidirectional triplet—the association pair (s, t) remains significant ($P < 0.05$) even upon conditioning on the transcript level t', but not vice versa. (ii) unidirectional triplet (s,t') remains significant even upon conditioning on the transcript level t, but not vice versa. (iii) bi-directional triplet (s,t) remains significant even upon conditioning on the transcript level t' and (s, t') remains significant even upon conditioning on the transcript level t. (iv) dependent triplet (s, t) and (s, t') are insignificant ($P > 0.05$) when conditioning on the transcript level t' and t respectively. (**b**) Graphical representation of intra-module interactions. We consider a module with three transcripts: t_1, t_2, t_3 and a main SNP s. A bi-directional dashed purple edge is placed between transcripts t_1 and t_2, representing the mutually independent association of both t_1 and t_2 with the main SNP s. A directed solid pink edge is placed between transcripts t_2 and t_3, representing the dependent association of (s,t_3) on the transcript levels of t_2. No edge is placed between transcripts t_1 and t_3, representing the mutually dependent association of both t_1 and t_3 with the main SNP s.

connects transcripts with directionally/mutually independent association with the main SNP (Figure 1b).

Distribution of gene expression levels across samples

We perform a Jarque–Bera test of the null hypothesis that the gene expression across samples comes from a normal distribution. Most of the genes did not comply with a normal distribution, but the percentage of genes inconsistent with normality was higher in pairs that reoccur in two permutations or more. This is a likely cause for the distorted statistics of these pairs. Although some previous studies take an approach of normalizing the data prior to analysis, we chose to rely on raw data and discard extreme results (see 'Results' section and Supplementary Text S2).

Module annotation

The enrichment of a module in gene subsets from the Gene Ontology (GO) (20), and KEGG (21) databases was calculated using DAVID (22,23). The enrichment of real and permuted modules in gene subsets from the NCBI gene database was calculated using LitVAn (19). We report only modules with annotations that have a significant FDR of 0.05 or better. Depending on context, we discuss the proximity of a gene to a SNP in several ways: A SNP may be 'in the span of the gene', i.e. the SNP resides between the ENSEMBL (24) transcription start site and stop codon of the gene; 'closest to the gene', i.e. this gene spans the closest among all spanned sites on either direction; or 'close to the gene'—means the SNP is within 1 Mb of a site spanned by the gene. We define a *cis* main SNP when the main SNP is within 1 Mb of one or more transcripts in the module. We define a *trans* main SNP when the main SNP is 1 Mb or further of all the transcripts in the module.

Enrichment of cis-effects for main SNPs

We model the examination of *cis*-effects for main SNPs as a binomial experiment. For each main SNP, we record one closest gene. Conservatively, unique genes are tested for association to exactly one main SNP, a binomial experiment Bin(n = number of unique genes, $P = 0.05$) with significant number of successes.

We then record main SNPs that are at least 1 Mb apart from one another and test them for association to exactly one closest gene, a binomial experiment Bin(n = number of main SNPs that are at least 1 Mb apart from one another, $P = 0.05$) with significant number of successes (Supplementary Table S3 and 'Results' section for full details).

Comparison with alternative method for module construction

We implemented the standard approach of grouping genes according to their associated SNP. We used a standard, stricter FDR cutoff of 10% for association–pairs (11). We compare these results with those reported by our own method (see 'Results' section).

RESULTS

Computational framework for detecting transcriptional modules

We set out to develop a statistical–computational framework to elucidate the regulatory structure by which genetic variants affect transcription. Specifically, we aim to examine the hypothesis that SNPs can have a modular effect on gene expression. Our method detects transcriptional modules, each including transcripts that are associated with the same main SNP. It is important to distinguish the modules that we find from co-expression clusters. Specifically, we represent each module as a graph, where nodes are transcripts, and for each possible pair of transcripts an edge correspond to a scenario where at least one of the transcripts remains significantly associated to the SNP when conditioned on its counterpart.

An initial step of detecting association pairs of SNP and transcript, showed as many such pairs as expected under the null hypothesis of no such true association. However, we were still motivated to search for modules, as the same associated SNPs were shared by many transcripts. Briefly, we collated association pairs that share a SNP into triplets and larger modules. Such modules are more numerous, bigger, denser in association and more functionally enriched than expected by chance.

In detail, we devised a three-step procedure for detecting the modules regulated by eQTLs.

The first step detects 67 540 association pairs of a SNP s and a transcript t whose expression level is putatively associated with s (nominal association $P < 10^{-5}$, see 'Materials and Methods' section for details). The distribution of the number of pairs in the permuted data (Figure 2a) demonstrates that the observed number of association pairs is consistent with the null expectation ($P \approx 0.07$). We eliminate 623 pairs that include transcripts whose association statistic is strongly distorted, as observed by permutation (see 'Materials and Methods' section for details). We proceed with analyzing the remaining 66 917 association pairs.

Association pairs are binned by SNP s, and give rise to 10 354 modules (see 'Materials and Methods' section), ranging in size from 2 to 91 transcripts (Supplementary Figure S2 and Supplementary Table S2) who are associated to the same main SNP of the module. Only 518 modules are large, i.e. with 10 or more transcripts. There are significantly more modules—10 354 (Figure 2b) than those found in the permuted data (average 2322 across permutations; SD 208; see 'Materials and Methods' section). Specifically, there are significantly more large modules—518 (Figure 2c) than those found in the permuted data (average 220; SD 42). We note the occurrence of one outlier permuted data set (see Supplementary Text S3 for details). While the observed number of significantly associated pairs of transcript and SNP is consistent with the null expectation, we find that there are significantly more modules than those found in the permuted data. This finding is consistent with the premise that gene regulation is modularly organized.

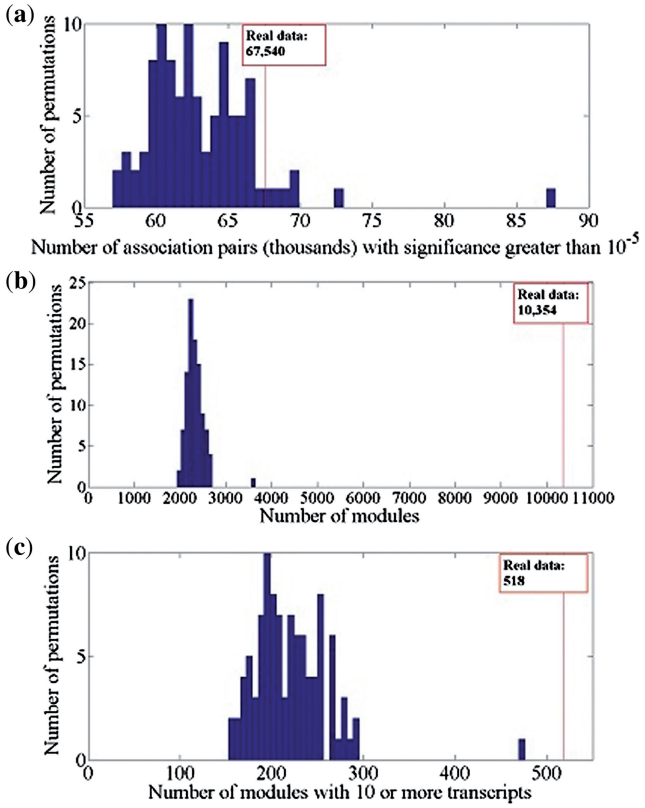

Figure 2. The number of (**a**) association pairs (**b**) modules and (**c**) large modules in real data compared with 100 permuted data sets. Although only 93 out of the 100 permutated data sets have fewer association pairs than in the real data, all of them have fewer (large) modules.

Modules' topology

The set of pairs includes 137 889 possible triplets (s,t,t') where (s,t) and (s,t') are association pairs. Focusing on co-associated pairs of transcripts, we find that for 129 130 of these triplets, association for at least one of the pairs, (s,t) remains significant ($P < 0.05$) even upon conditioning on the transcript level of t' (see 'Materials and Methods' section). These triplets are further subdivided into the 101 762 'bi-directional' triplets versus the remaining 27 368 'uni-directional' (for definitions see 'Materials and Methods' section).

We describe independence of associations in each module M as a graph $G(M)$ (see 'Materials and Methods' section), when examining the topology of the modules, we notice that for most modules, nearly all association pairs are mutually independent (Figure 3 and Supplementary Figure S3). Furthermore, considering all possible pairs of transcripts in a module, the fraction of them which were connected by edges is 87.7% (averaged across all modules; SD 13.3%). This is significantly more than those found in permuted data (average 12.5%;SD 6.2%). Specifically, both bi-directional (average 79.4%, SD 18.9% versus average 2.3%;SD 3.1%), as well as uni-directional edges (average 8.3%; SD 6.3% versus 10.2%; SD 5.2%) are enriched in real compared with permuted data (Supplementary Figure S3). This is

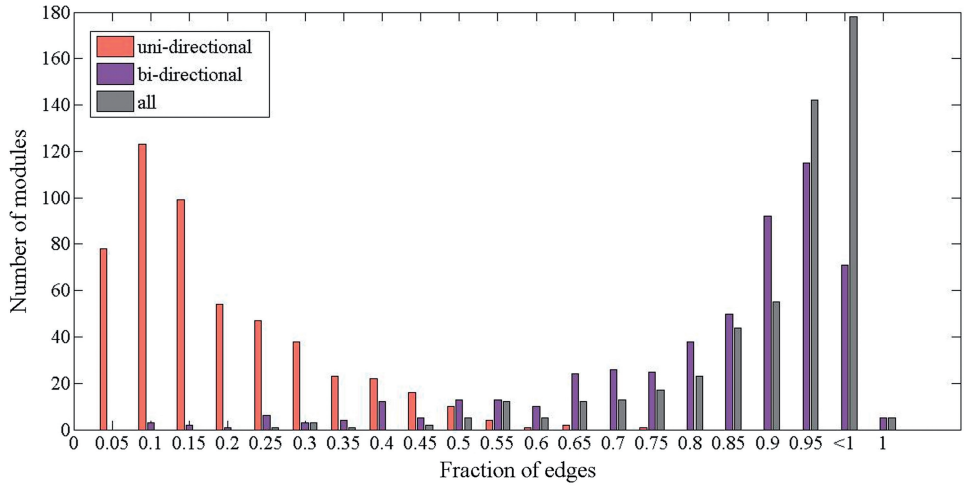

Figure 3. The distribution of the number of modules with different fractions of edges This figure shows the distribution of the number of modules with different fractions of uni, bi- and all edges represented in pink, purple and gray, respectively in each one of the 518 large modules.

consistent with the main SNP affecting expression levels of most transcripts in its module in a simultaneous rather than a cascaded manner. This also addresses concerns of artifactual modules that are possibly just clusters of co-expressed genes rather than truly independent association to the main SNP.

Module's score and filtering

To establish a measure of confidence in the resulting modules, we assign a score to each module, considering the module size and the strength of associations between the main SNP and each of the transcripts in the module. This score is justified as a log-likelihood-ratio that compares two hypotheses (see 'Materials and Methods' section). We provide an empirical *P*-value interpretation by scaling the scores of modules in the real data, compared with the average score of the modules in permutations. We further prune the large modules, defining a subset of 114 high confidence modules with FDR < 0.02 (Supplementary Figure S4).

We notice that in most of the modules there are few transcripts that are expressed in an opposite direction to the majority of transcripts in the module. This suggests that the main SNP affects the majority of transcripts in the same direction. We verify this observation by quantifying the percentage of positive and negative correlation of the main SNP with the transcripts in each module (Supplementary Figure S5).

Cis/*trans*-effects

Some of the previous studies have optimized power to detect *cis*-regulatory variation by using different *P*-value threshold for defining *cis* eSNPs (7), based on strong priors in their favor (10). Here, we set a fixed threshold of 10^{-5} for both *cis*, and *trans* association, putting them on equal footing for the detection of modules.

There are 110 modules with *trans* main SNP, the remaining 4 modules have *cis* main SNP (See 'Materials and Methods' section for definitions). We systematically sought potential *cis*-effects of main SNPs that were not strong enough to be captured by our first-pass analysis. To examine this, we record the gene closest (see 'Materials and Methods' section) to each main SNP. In two modules, the main SNPs did not have a close gene from our data. The main SNPs of the remaining 112 modules have 94 unique closest genes, which we call 'main genes'. Out of all main SNPs, 88 are at least 1 Mb apart from one another. More details on grouping the main SNPs according to different categories can be found in the supplementary material (Supplementary Text S4; Supplementary Tables S3 and S4). We record the *P*-value for the linear regression between each main SNP and the expression levels of its closest gene. In total, 24 main SNPs were nominally ($P < 0.05$) *cis* associated to their respective closest gene, with 14 unique associated genes ($P = 1.76 \times 10^{-4}$, see 'Materials and Methods' section) and with 10 unique associated SNPs that are at least 1 Mb apart from one another ($P = 8.1 \times 10^{-3}$, see 'Materials and Methods' section). These main SNPs are *trans* main SNPs. These results support our suggested *trans*-effect model.

Independent cross validation by similar annotations from two sources of information and phenotypic analysis

We characterized high confidence modules by considering two sources of information:

(i) the enrichment of transcripts in a module for membership in gene-sets from the Gene Ontology (20), NCBI Gene and KEGG (21) databases.
Of the 114 modules, 26 (22.8%) were reported as enriched in any category. This contrasts with modules in 100 permuted data sets, where $12.8 \pm 2.7\%$ of the modules show any functional enrichment (Supplementary Figure S6) and

(ii) locus annotation of the main and secondary SNPs of each module, as reflected in the existing literature, Ensembl (24) and wikigenes (25).

These sources are independent for modules with *trans* main SNP. We observe similar annotations of modules from the two sources of information. This independent cross validation provides support for our methodology. Additional support comes from intersecting the 94 main loci with the 2626 unique genes (2212 among the 18 873 transcripts available for analysis in this work) reported to house GWAS SNPs (26). We find an overlap of 21 genes (hypergeometric $P = 1.1 \times 10^{-3}$).

We discard 19 modules whose set of transcripts have a ≥90% overlap with other modules, resulting in 95 distinct modules (see Supplementary Text S4; Supplementary Tables S4, S5a and S5b for full listing of all 95 modules). We present details of the annotation analysis for three modules: the largest with an annotated *cis*-SNP, and two of the four largest modules overall.

Comparison with standard approach to module construction

We show the standard approach (see 'Materials and Methods' section) to produce fewer modules, smaller modules, limiting its use for finding modules. Moreover, our approach finds modules that are more enriched for functional annotation categories, compared with the standard approach, supporting our modules being genuine.

Specifically, the standard approach produced 22 015 association pairs, 3387 modules, 75 with 10 transcripts or more (Supplementary Figure S7). The largest module has 27 transcripts. We examine the enrichment of these modules in GO categories and KEGG pathways: 4 out of the 75 modules had significant biological enrichment in at least one category (5.3% comparing with 22.8% functional enrichment in our modules).

Support for modules filtering step

All four modules that were found by the standard method and were functionally enriched are contained in one of our final 95 modules. This provides a support for our module scoring and filtering step.

Analysis of specific modules

We present a positive control for our method using module no. 29 with 16 transcripts and *cis* main SNP. The main SNP rs9267658 partitions the samples into three groups: 277 samples that are homozygous C (C/C), 89 C/T samples and 5 T/T samples. The secondary SNP for the C/T subgroup of samples is rs4902609 and is associated with eight transcripts. This module is enriched for Major histocompatibility Complex (MHC) genes (FDR 0.0049), with related annotation for relevant KEGG pathways (allograft rejection—FDR 0.0046, antigen processing and presentation—FDR 0.0041, cell adhesion molecules—FDR 0.0088) and autoimmune diseases (graft-versus-host disease—FDR 0.0027, type I diabetes mellitus—FDR 0.0021, thyroid disease—FDR 0.0023, viral myocarditis—FDR 0.0036 and asthma—FDR 0.045). The main SNP resides within the MHC region (27). The module includes three transcripts in *cis* to the main SNP that play a central role in the immune system: *HLA-DRB5*, *HLA-DRB4* are MHC class II and *HLA-G* is MHC class I. The closest gene to rs4902609, *RAD51L1* is a tumor suppressor gene, whose *trans*-association to the MHC transcripts may relate to previous reports on links between autoimmunity and cancer (28) (Figure 4).

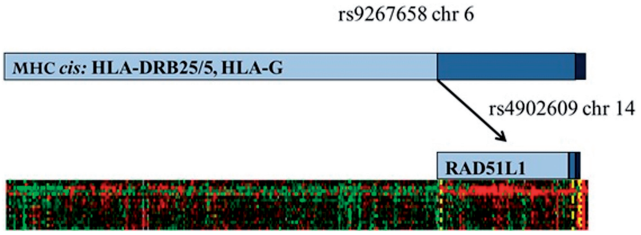

Figure 4. Module of size 16 transcripts and their expression levels over 371 samples. The heatmap of expression levels (red/black/green) across samples (columns) and genes (rows) is segmented (top) into SNP–genotype splits—light, medium and dark blue represent carriers of 0, 1 or 2 minor alleles, respectively. Closest genes to the main and secondary SNP are listed.

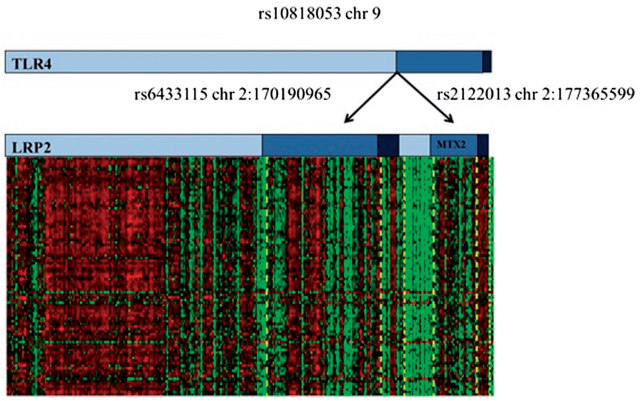

Figure 5. The largest module of 91 transcripts and their expression levels over 371 samples (Figure 4 legend for further details).

The largest module (#1) has 91 transcripts. The main SNP rs10818053 partitions the samples into 303 T/T samples, 65 T/C samples and 3 C/C samples. The secondary SNPs are rs6433115 for the major-allele homozygotes and rs2122013 for heterozygotes. This module is enriched for transcripts involved in oxidation reduction (FDR 5.9×10^{-15}), lipid metabolic processes (FDR 1.9×10^{-5}) and genes expressed in the mitochondrion (FDR 0.015). In terms of pathways, it is enriched for drug metabolism pathways (FDR 7.7×10^{-5}) and primary bile acid biosynthesis (FDR 3.4×10^{-4}) that occurs in the liver. The closest gene to the main SNP, *TLR4* cooperates to mediate the innate immune response to bacterial lipopolysaccharide (LPS). *TLR4* activation mediates liver inflammatory response (29) and is responsible for oxidized phospholipid-mediated inhibition of *TLR* signaling (30). Secondary SNP rs6433115 for the T/T subgroup is associated with 26 transcripts and is within the span of *LRP2*. Secondary SNP rs2122013 for the T/C subgroup is associated with 35 transcripts and is closest to *MTX2* gene (Figure 5). *LRP2* is a lipoprotein that is also involved in

Table 1. Data for liver risk in 371 samples separated by minor–minor, major–minor and major–major allele samples, respectively and genotype of rs10818053

rs10818053 genotype	Liver risk			
	Minor–minor C/C	Major–minor T/C	Major–major T/T	Total no. of samples
Positive	2	13	39	54
Negative	1	52	264	317
Total no. of samples	3	65	303	371

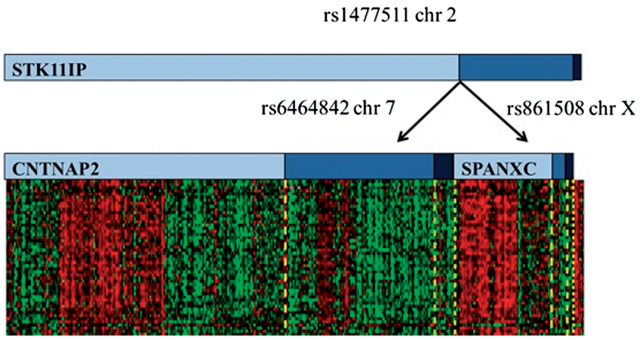

Figure 6. Module of size 50 transcripts and their expression levels over 371 samples. (Figure 4 legend for further details).

the cellular uptake of drugs, including lipid-based formulations (31). *MTX2* is involved in the import of proteins into the mitochondrion (25). This module may be related to the effect of drugs on lipid metabolism (32) and the possible role of the mitochondrion in such pathways (33).

Since mutations in *TLR4* are associated with liver damage, we investigate the main SNP's association to drug sensitivity. Data for liver risk in the 371 samples (11), genotype of the main SNP rs10818053 and liver risk in 371 samples are detailed in Table 1. The clinical data presented by Schadt *et al.* (11) for liver risk, are binary entries describing (according to clinicians' diagnosis) if there is a risk to the patient's liver if treated by drugs. We present preliminary analysis showing that these minor–minor and major–minor allele samples are enriched for liver risk more than is expected by chance (Hypergeometric $P < 0.012$) which implies that individuals carrying C/C or T/C alleles in the main SNP's locus may be prone to liver sensitivity for drug treatment. This analysis provides the first support for our method from non-expression traits.

Module #4 has 50 transcripts. The main SNP rs1477511 partitions the samples into 288 T/T samples, 76 T/G samples and 7 G/G samples. The secondary SNPs are rs6464842 for the first subgroup and rs861508 for the second subgroup and are associated with 7 and 9 transcripts, respectively (Figure 6). This module is enriched in transcripts that regulate cellular (FDR 0.0036) and metabolic processes (FDR 0.013), specifically cell proliferation and differentiation (FDR 5.2×10^{-5}). It is enriched for ErbB (FDR 1.5×10^{-3}) and Mitogen activated protein kinase (MAPK) signaling pathways (FDR 5.2×10^{-3}). The closest gene to the main SNP, *STK11IP* interacts with *LKB1* which regulates cell polarity and functions as a tumor suppressor (25). *LKB1* is a serine/threonine kinase which is inactivated by mutation in the Peutz–Jeghers polyposis and cancer predisposition syndrome (PJS) (34), with correlation to the putative function of the module. We observe a significant *P*-value (<0.031) between the expression levels of *LKB1* and the genotype of rs1477511. Mutations in *CNTNAP2*, where rs6464842 resides, have been implicated in multiple neurodevelopmental disorders, including attention deficit hyperactivity disorder (ADHD) and schizophrenia. With correlation to *CNTNAP2* function, the ErbB signaling was suggested to impair working memory and executive functions that are affected in schizophrenia, ADHD and other psychiatric disorders (35). *SPANXC* which is the closest gene to rs861508 resides in a region that confers susceptibility to prostate cancer. ErbB and MAPK signaling are known to have an important role in cancer (36,37).

Table 2. Data for alcohol risk in the 371 samples, genotype of rs1453226 and alcohol risk in minor–minor allele samples

rs1453226 genotype	Alcohol risk		
	Minor–minor A/A	Major–major G/G and Major–minor G/A	Total no. of samples
Positive	4	15	19
Negative	4	93	97
Unknown	28	227	255
Total no. of samples	36	335–195 and 140, respectively	371

Finally, we present a second support for our method from non-expression traits. Module #101 with 10 transcripts is the only module where the main SNP maps to a locus associated with oxidative damage control: rs1453226 at *OXR1* indicated to be involved in protection from oxidative damage (25). The transcripts in this module are slightly enriched for oxoacid metabolic process (FDR 0.04). Therefore, we decided to investigate its association to alcohol risk. Data for alcohol risk in the 371 samples (11), genotype of the main SNP rs1453226 and alcohol risk in minor–minor allele samples are detailed in Table 2. It is challenging to provide clinical support, since the clinical data presented by Schadt *et al.* (11) is very sparse. We present preliminary analysis showing that these samples are enriched for alcohol risk more than is expected by chance (Hypergeometric $P < 0.03483$), which implies that individuals carrying A/A alleles in the main SNP's locus may be prone to sensitivity for alcohol use.

DISCUSSION

We presented a three-step approach to the analysis of eSNPs and their relation to phenotypes that goes beyond documenting associations of each to expression levels, by applying a module score filtering procedure,

and complements co-expression networks by unraveling module topology. As a first step, we assemble transcripts associated to the same main eSNP into the modules. We then filter the reported modules by a confidence score, and finally associate subgroups of transcripts within a module with additional variants conditioned on the genotype of the main SNP.

We apply our method to data on human liver expression and SNP genotypes (11). We find that the number of association pairs of eSNP and transcript is consistent with the null expectation, whereas assembled modules are significantly more numerous, bigger and denser than those observed in the permuted data. This indicates modules are not random clusters of correlated-expression genes, but rather show truly independent association to their main SNP. We compare our results with a standard approach that maps transcript-eQTL pairs with a standard FDR (e.g. 10%) and forms groups consisting of transcripts that share an eQTL. We observe smaller number of modules, smaller in size and significantly less enriched in Biological categories.

Our method detects 95 distinct modules; out of those, only one has a main SNP in *cis* to module transcripts. Among the remaining 94 *trans* main eSNPs, we observe enrichment for milder, not genome-wide significant *cis*-effects that explain the *trans*-effect of the main SNPs on transcripts in the associated modules. We characterize modules by two sources of information that are independent for modules with a *trans* main SNP: enrichment in subsets of genes and locus annotation of the main and secondary SNPs. We observe similar annotations from both sources of information. Thus, providing support for our method. We present detailed analysis of four modules: annotation analysis for three of the four modules: one with a *cis* main SNP and two with *trans* main SNPs, and phenotypic analysis for two of the four modules.

This study holds the promise for extension beyond its current limitations. The current analysis focuses on transcripts that are directly regulated by a variant. Mining the data for additional transcripts that are downstream along the same pathway of regulation, e.g. by consideration of co-expressed genes with milder association to the main SNP can complement reverse engineering of the regulatory program (8). Furthermore, both the raw data sets (11) and supporting databases (20,21,24) in this work are noisy and limited. Potential increase in sample size for eQTL data may enable detection of eSNP associations at more significant *P*-values for even milder effects. Likewise, as the functional annotation continues to build up, better understanding of modules would be facilitated.

Future studies could extend the approach presented here to investigate how modules correlate with phenotype, for example, using the data on enzymatic activity that was presented by Yang *et al.* (14). As data becomes available, comparison of modular structure between healthy and affected samples, as well as across different tissue types is likely to improve understanding of disease and developmental regulatory processes. It remains a significant challenge to validate the results presented here by experimental means, and analysis of independent data may provide such validation by replication.

AVAILABILITY

The method presented in this work is publicly available at http://www.cs.columbia.edu/~itsik/IMR_eQTL2/Inference_of_modules_associated_to_eQTLs.htm.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–5; Supplementary Figures 1–7; Supplementary References [24,25] and Supplementary Text 1–4.

FUNDING

Case Western – *5-23635* CWRU RES503641 [National Institute of Health]; MAGNET – *5-76642* 2U54CA121852-07 SC5; NSF CAREER – *5-24527* IIS-0845677; ELLIPSE – *5-20049* USC H50431.

Conflict of interest statement. None declared.

REFERENCES

1. Cheung,V.G. and Spielman,R.S. (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat. Rev. Genet.*, **10**, 595–604.
2. 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.
3. Rockman,M.V. and Kruglyak,L. (2006) Genetics of global gene expression. *Nat. Rev. Genet.*, **7**, 862–872.
4. 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.
5. Moffatt,M.F., Kabesch,M., Liang,L., Dixon,A.L., Strachan,D., Heath,S., Depner,M., von Berg,A., Bufe,A., Rietschel,E. *et al.* (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, **448**, 470–473.
6. Kathiresan,S., Melander,O., Guiducci,C., Surti,A., Burtt,N.P., Rieder,M.J., Cooper,G.M., Roos,C., Voight,B.F., Havulinna,A.S. *et al.* (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.*, **40**, 189–197.
7. 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.
8. Litvin,O., Causton,H.C., Chen,B.J. and Pe'er,D. (2009) Modularity and interactions in the genetics of gene expression. *Proc. Natl Acad. Sci. USA*, **106**, 6441–6446.
9. Gilad,Y., Rifkin,S.A. and Pritchard,J.K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, **24**, 408–415.
10. Stranger,B.E., Forrest,M.S., Clark,A.G., Minichiello,M.J., Deutsch,S., Lyle,R., Hunt,S., Kahl,B., Antonarakis,S.E., Tavare,S. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.
11. 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.
12. Ghazalpour,A., Doss,S., Zhang,B., Wang,S., Plaisier,C., Castellanos,R., Brozell,A., Schadt,E.E., Drake,T.A., Lusis,A.J. *et al.* (2006) Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.*, **2**, e130.
13. 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.
14. 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.
15. Hartwell,L.H., Hopfield,J.J., Leibler,S. and Murray,A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.
16. Ihmels,J., Bergmann,S., Berman,J. and Barkai,N. (2005) Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program. *PLoS Genet.*, **1**, e39.
17. 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.
18. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., de Bakker,P.I., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
19. Akavia,U.D., Litvin,O., Kim,J., Sanchez-Garcia,F., Kotliar,D., Causton,H.C., Pochanard,P., Mozes,E., Garraway,L.A. and Pe'er,D. (2010) An integrated approach to uncover drivers of cancer. *Cell*, **143**, 1005–1017.
20. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
21. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
22. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
23. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
24. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.
25. Hoffmann,R. (2008) A wiki for the life sciences where authorship matters. *Nat. Genet.*, **40**, 1047–1051.
26. 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.
27. Lee,H.S., Lee,A.T., Criswell,L.A., Seldin,M.F., Amos,C.I., Carulli,J.P., Navarrete,C., Remmers,E.F., Kastner,D.L., Plenge,R.M. *et al.* (2008) Several regions in the major histocompatibility complex confer risk for anti-CCP-antibody positive rheumatoid arthritis, independent of the DRB1 locus. *Mol. Med.*, **14**, 293–300.
28. Mantovani,A., Allavena,P., Sica,A. and Balkwill,F. (2008) Cancer-related inflammation. *Nature*, **454**, 436–444.
29. Zhai,Y., Shen,X.D., O'Connell,R., Gao,F., Lassman,C., Busuttil,R.W., Cheng,G. and Kupiec-Weglinski,J.W. (2004) Cutting edge: TLR4 activation mediates liver ischemia/reperfusion inflammatory response via IFN regulatory factor 3-dependent MyD88-independent pathway. *J. Immunol.*, **173**, 7115–7119.
30. Erridge,C., Kennedy,S., Spickett,C.M. and Webb,D.J. (2008) Oxidized phospholipid inhibition of toll-like receptor (TLR) signaling is restricted to TLR2 and TLR4: roles for CD14, LPS-binding protein, and MD2 as targets for specificity of inhibition. *J. Biol. Chem.*, **283**, 24748–24759.
31. Wasan,K.M., Brocks,D.R., Lee,S.D., Sachs-Barrable,K. and Thornton,S.J. (2008) Impact of lipoproteins on the biological activity and disposition of hydrophobic drugs: implications for drug discovery. *Nat. Rev. Drug Discov.*, **7**, 84–99.
32. Thomas,E.A., George,R.C., Danielson,P.E., Nelson,P.A., Warren,A.J., Lo,D. and Sutcliffe,J.G. (2003) Antipsychotic drug treatment alters expression of mRNAs encoding lipid metabolism-related proteins. *Mol. Psychiatry*, **8**, 983–993.
33. Canto,C., Gerhart-Hines,Z., Feige,J.N., Lagouge,M., Noriega,L., Milne,J.C., Elliott,P.J., Puigserver,P. and Auwerx,J. (2009) AMPK regulates energy expenditure by modulating NAD+ metabolism and SIRT1 activity. *Nature*, **458**, 1056–1060.
34. Smith,D.P., Rayter,S.I., Niederlander,C., Spicer,J., Jones,C.M. and Ashworth,A. (2001) LIP1, a cytoplasmic protein functionally linked to the Peutz-Jeghers syndrome kinase LKB1. *Hum. Mol. Genet.*, **10**, 2869–2877.
35. Kwon,O.B., Paredes,D., Gonzalez,C.M., Neddens,J., Hernandez,L., Vullhorst,D. and Buonanno,A. (2008) Neuregulin-1 regulates LTP at CA1 hippocampal synapses through activation of dopamine D4 receptors. *Proc. Natl Acad. Sci. USA*, **105**, 15587–15592.
36. Dhillon,A.S., Hagan,S., Rath,O. and Kolch,W. (2007) MAP kinase signalling pathways in cancer. *Oncogene*, **26**, 3279–3290.
37. Hynes,N.E. and Lane,H.A. (2005) ERBB receptors and cancer: the complexity of targeted inhibitors. *Nat. Rev. Cancer*, **5**, 341–354.