Human Molecular Genetics, 2014, Vol. 23, No. 10 **2780–2790**
doi:10.1093/hmg/ddt668
Advance Access published on December 30, 2013

Guilt by rewiring: gene prioritization through network rewiring in Genome Wide Association Studies

Lin Hou[1], Min Chen[2], Clarence K. Zhang[3], Judy Cho[4] and Hongyu Zhao[1,*]

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA, [2]Quantitative Biomedical Research Center, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA, [3]Keck Biotechnology Laboratory Biostatistics Resource, Yale School of Medicine, New Haven, CT 06510, USA and [4]Internal Medicine, Yale School of Medicine, New Haven, CT 06510, USA

Received July 14, 2013; Revised December 5, 2013; Accepted December 23, 2013

Although Genome Wide Association Studies (GWAS) have identified many susceptibility loci for common diseases, they only explain a small portion of heritability. It is challenging to identify the remaining disease loci because their association signals are likely weak and difficult to identify among millions of candidates. One potentially useful direction to increase statistical power is to incorporate functional genomics information, especially gene expression networks, to prioritize GWAS signals. Most current methods utilizing network information to prioritize disease genes are based on the 'guilt by association' principle, in which networks are treated as static, and disease-associated genes are assumed to locate closer with each other than random pairs in the network. In contrast, we propose a novel 'guilt by rewiring' principle. Studying the dynamics of gene networks between controls and patients, this principle assumes that disease genes more likely undergo rewiring in patients, whereas most of the network remains unaffected in disease condition. To demonstrate this principle, we consider the changes of co-expression networks in Crohn's disease patients and controls, and how network dynamics reveals information on disease associations. Our results demonstrate that network rewiring is abundant in the immune system, and disease-associated genes are more likely to be rewired in patients. To integrate this network rewiring feature and GWAS signals, we propose to use the Markov random field framework to integrate network information to prioritize genes. Applications in Crohn's disease and Parkinson's disease show that this framework leads to more replicable results, and implicates potentially disease-associated pathways.

INTRODUCTION

Genome Wide Association Studies (GWAS) have uncovered many susceptible loci underlying common genetic disorders, and provided new insights into disease aetiology (1). Since the whole genome is scanned to identify signals in GWAS, many novel pathways/genes have been found to affect disease risk. For example, 140 loci have been found associated with Crohn's disease (CD) (2,3), including many belonging to pathways not suspected to be involved in CD before. In spite of these successes, much more research is needed to most comprehensively analyse and extract information from these rich data. First, it is believed that many loci remain to be discovered, because association signals are often weak and difficult to be separated from background noise (4). These undiscovered loci are likely to be as important as those already identified to understand disease aetiology and identify treatment strategies. Second, the replication rate between studies is low, especially for those markers with marginal statistical evidence of association. Third, although the hypothesis free nature of GWAS has led to many unexpected discoveries, it is challenging to relate the associated markers with disease mechanism (5). Thus, there is a great need to develop gene prioritization methods that generate more replicable results and biologically plausible candidates by incorporation of biological prior information.

A number of computational frameworks have been proposed to prioritize candidate genes by incorporating functional genomics data, including linkage analysis results, gene annotations, sequence features, eQTLs, protein interaction networks and biological pathways (6–8). These methods have been shown to have advantages in identifying SNPs/genes with increased biological relevance, and enriching signals in GWAS of the

[*]To whom correspondence should be addressed at: Hongyu Zhao, Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA. Tel: +1 2037853613; Fax: +1 2037856912; Email: hongyu.zhao@yale.edu

disease. However, the replication rate of prioritization results between independent studies has not been thoroughly evaluated. In this manuscript, our primary interest is to integrate network information in post-GWAS prioritization of candidate genes. In networks, nodes represent genes or proteins, and edges can be physical protein interactions, gene co-expressions (which can be inferred), or functional interactions derived from computational models (9,10). Regardless of the network used, all existing network-based approaches are built on the 'guilt by association' principle (11,12), which assumes that the state of association can be propagated through connections in the network. In other words, genes that are within close proximity with a disease gene in the network are likely related to the disease themselves as well. Various algorithms have been developed to predict or prioritize disease genes by this 'guilt by association' approach (13–15). Chen *et al.* used page-rank algorithm to prioritize candidate genes in a protein interaction network (6); Lee *et al.* (10) built a functional gene interaction network, and used a label propagation algorithm to prioritize candidate genes. In their paper, the overall functional interaction was defined by a weighted sum of log likelihood scores, with each score derived from the naïve Bayesian classifier trained for one of the four data sources: mRNA expression, protein–protein interaction, protein complex, and comparative genomics.

This 'guilt by association' approach has proved useful in gene prioritizations. Yet, the major drawback of this line of work is that a network is treated as a static reference, which does not reflect the dynamic nature of biological networks. In yeast, for instance, interrogating the transcriptional regulatory network under multiple conditions uncovered extensive rewiring of the network architecture (16). Indeed, cellular networks are frequently rewired to respond to different stimuli, and valuable insight can be gained by studying networks from a dynamics perspective (17). Bandyopadhyay *et al.* (18) reported experiments that studied the genetic interaction network of the budding yeast in normal conditions and that in DNA damage stimuli. Novel gene functions and links in DNA repair pathways were uncovered by contrasting these two networks, which could not be detected by looking at static networks alone. Particularly to our interest in complex disease genetics, network changes in disease samples might provide clues of disease aetiology. However, the study of network dynamics in higher eukaryotes is hindered by the complexity of the genome, which makes genome scale interrogation of interactions through experimental approaches at multiple conditions extremely labour-intensive and expensive. Alternatively, we can infer co-expression network, an indirect interaction network, from mRNA expression data sets in both patient and control samples. Microarray experiments are robust and affordable, and many data sets are available in public databases like GEO (19) for many diseases, thus providing a rich data source to investigate the dynamic perspective of networks. To identify network components that differ between the affected and healthy individuals, co-expression networks in disease samples and control samples can be constructed separately through co-expression analysis and contrasted to identify the changing elements between the networks. For example, by investigating differential wiring between transcription factors and target genes in two groups of cattle crosses, researchers were able to pinpoint a DNA mutation in myostatin as the causal mutation for muscle change in cattle (20). Another example was the study by Taylor *et al.* (21) who used changes of Pearson correlation coefficients (PCC) between genes to study the dynamic structure of protein interaction networks, and found alterations in modularity are related with breast cancer prognosis.

In this study, we propose a 'guilt by rewiring' approach, which explores the dynamic aspect of the network, i.e. network rewiring, by comparing the gene co-expression networks in disease samples and control samples. Using CD as a proof of principle example, we demonstrate that disease-related genes are more frequently rewired as opposed to a random gene in the genome. Furthermore, we propose a Markov random field model to incorporate network-rewiring information to prioritize GWAS signals. The applications of this approach to CD and Parkinson's disease (PD) show that our approach generates more replicable results. Moreover, the proposed method identified known pathways associated with these diseases, and also implicated potential pathways that have not been suggested by existing methods.

RESULTS

Network rewiring is abundant in disease-related pathways

We use CD as an example to demonstrate that network rewiring occurs more frequently for disease-related genes than other genes in the whole genome. In our analysis, the whole genome refers to the intersection set of genes in the GWAS study and the microarray study of CD, consisting of 12 007 genes (see Supplementary Material, Table S2).

CD is an immune-mediated disease, thus we investigated the frequency of rewiring within genes in the immune processes, and that in the whole genome. The CD microarray data set (GEO accession number: GSE20881) profiled mRNA expressions of 172 intestine biopsies from CD patients and control subjects. To measure network rewiring, gene co-expression networks in CD samples and control samples were constructed separately by calculating the pairwise PCC of gene expression profiles. A pair of genes was considered to be rewired if they were co-expressed under one condition, and not under the other condition. More specifically, we used Fisher's method to test the significance of the difference between PCCs (see Materials and Methods), and the degree of rewiring was defined as one minus the *P*-value of the test. Thus, the degree of rewiring ranges between 0 and 1, and a larger value indicates a more significant change of co-expression between disease and control conditions. The rewiring network is a weighted network, and edges are weighted by the degree of rewiring between disease and healthy conditions. Several thresholds were considered (the thresholds were chosen so that the density of the discrete network was 0.1, 0.05 and 0.01, see Supplementary Material, Table S3 for the corresponding thresholds) to dichotomize the rewiring network.

To investigate the density of rewiring edges in immune-related pathways, we defined the gene set of 'immune system' based on the Reactome annotation (22). In Reactome, 933 genes are documented in this category, and 603 genes remained after intersection with the genes that passed quality control in the microarray data set (see Materials and Methods). When we dichotomized the network at various thresholds, the density in the subnetwork of the immune system was consistently higher than that of the overall network (see Fig. 1). Thus, the immune system was significantly enriched with rewiring edges at varying thresholds (binomial test,

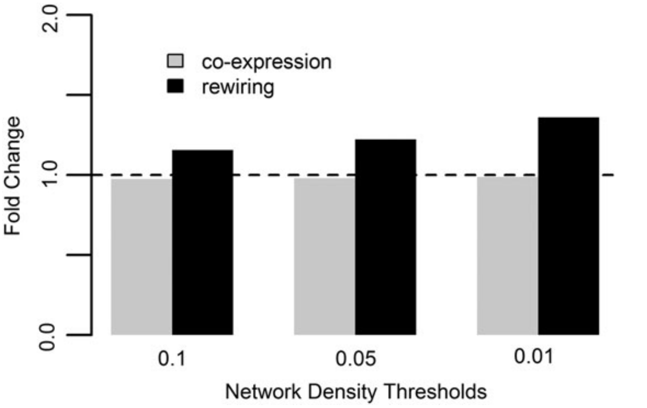

Figure 1. Network rewiring is abundant in immune system (defined in the Reactome pathway). The *y*-axis is the fold change of density in the subnetwork of the immune system to the total network. Network density is the number of edges in the observed network over the number of all possible edges.

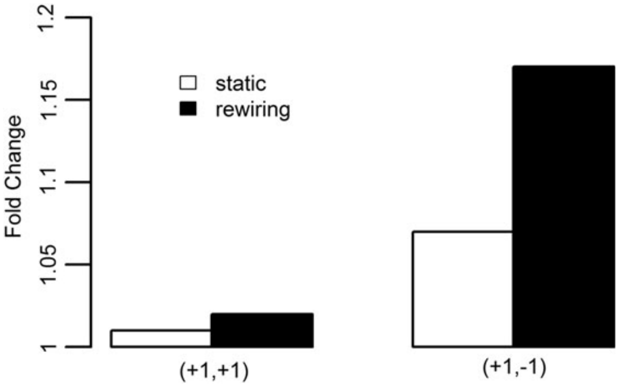

Figure 2. Comparisons of edge frequency in static network and rewiring network. (+1, +1) denotes an edge that connects two associated genes, while (+1, −1) denotes an edge that connects one associated gene and one non-associated gene. The *y*-axis is the fold change between observed number of edges and that of random expectation. Both the static and rewiring network were dichotomized at a threshold so that the corresponding network density is 0.01.

P-value ≤ 0.001). For example, the number of rewiring edges was 2465, as opposed to 1815 expected by chance, when the density of the rewiring network was 0.01. Interestingly, in the co-expression network defined by the healthy controls, the network density of the immune-related subnetwork was lower than that of the overall network.

Network rewiring is more informative than static co-expression network

To investigate the difference between the 'guilt by association' approach and 'guilt by rewiring' approach, we performed the following empirical analysis. A rewiring network, as described previously, and a static network was constructed separately for the same set of genes. To keep the two networks at comparable size after dichotomization, we varied the threshold for defining edges so that the two networks had the same network density (0.1, 0.05 and 0.01, respectively where the corresponding thresholds are shown in Supplementary Material, Table S3). In the static network, two genes were connected if they were co-expressed in the control samples (absolute PCC greater than the cut-off). Similarly in the rewiring network, two genes were connected if the degree of rewiring was greater than the cut-off. Genes were labelled, either 'associated' or 'not associated', based on the statistical evidence of their associations with CD, which was derived from the WTCCC GWAS study of CD (see Materials and Methods). A gene was labelled 'associated' (+1) if its association *P*-value was <0.05, and 'not associated' (−1) otherwise. Thus, in both networks, there were three types of edges, (+1, +1), (+1, −1) and (−1, −1). We posed the following question: which one of these two networks, static network or rewiring network, was more informative for association status? More precisely, were edges of types (+1, +1) and (+1, −1) enriched compared with that expected by chance? Since our proposed approach was motivated to prioritize candidate genes in GWAS instead of *de novo* prediction of disease genes, in all the analyses below (if not mentioned otherwise), the background genes are the set of genes with *P*-values <0.5 in the corresponding GWAS data set.

To answer this question, we calculated the proportion of associated genes, and the frequencies of the three types of edges in these two networks (Fig. 2). Under the null hypothesis, edge assignments are independent of association states of the two nodes, the proportions of edges (+1, +1), (−1, +1) and (−1, −1) should be a^2, $2a(1 - a)$ and $(1 - a)^2$, respectively, where a is the proportion of associated genes. For each network, we tested whether edges (+1, +1) and (+1, −1) were enriched with a binomial test (see Materials and Methods). When the network density was 0.01 (see Fig. 2), the fold change of (+1, +1) edges was 1.02 and 1.17 for the static and rewiring networks, respectively, both statistically more abundant than expected (*P*-values are 0). Similarly, the fold change of (+1, −1) edges was 1.01 and 1.07 for the static and rewiring networks (*P*-values are 0). A disease-associated gene might be affected in disease condition in different ways, such as expression level, phosphorylation status, structural conformation, so its interactions with the rest of the genome, regardless of the disease association status, would be subject to change. These results support the principles of both 'guilt by association' and 'guilt by rewiring'; however, the enrichment was more pronounced in the rewiring network. We also note that a modest increase of association signals in the rewiring network compared with the static network was consistently observed at various thresholds (see Supplementary Material, Table S4 for results at other thresholds).

Markov random field modelling approach

We adapted the Markov random field framework previously proposed by Chen *et al*. (23), which aimed at incorporating known pathway structure in GWAS analysis, to model the configuration distribution of the rewiring network. There are two advantages of the Markov random field model. First, the model can incorporate network structures, which account for long distance dependencies in associate states. Secondly, the computational framework through Markov chain Monte Carlo is well established. Basically, a Markov random field model is characterized by three elements: the state of each node in the network, the edges between nodes and the weights of the edges. In our framework, each node is a gene, and the state is either $+1$ or -1, indicating 'associated' or 'not associated'. Two genes are connected by an edge if they were co-expressed either in the disease samples or healthy samples. The edges are weighted by rewiring information. The model is specified to ensure that genes with more rewiring edges are more likely to be disease associated. Details of the model are provided in Materials and Methods.

Application to CD

We applied the proposed method to previously published CD GWAS. A good prioritization algorithm should generate results that are more replicable in independent cohorts. To test the performance of our method, we analysed one CD GWAS data set (denoted as NIDDK GWAS hereafter), and assessed the replication rate in a second one (denoted as WTCCC GWAS hereafter). In the NIDDK GWAS data set, the DNA samples were collected by the North American National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) IBD Genetics Consortium (IBDGC), and genotyped using Illumina HumanHap300 Genotyping BeadChip (24). The WTCCC CD data set was collected by The Wellcome Trust Control Consortium, and genotyped using Affymetrix GeneChip 500K. We used the cohort with fewer samples (NIDDK) to prioritize GWAS signals, and assessed the replication rate of the prioritized genes using the cohort with more samples (WTCCC GWAS).

Initially, there were 20 427 genes in the NIDDK GWAS, and 12 007 of them overlapped with qualified probe sets in CD microarray data. Only genes with association P-value <0.5 were included in the model, since our interest was to prioritize GWAS signals instead of *de novo* prediction, and the final number of genes included in the network was 5583. The posterior probability that a node is associated with CD can be calculated through Eq. (9). Genes can be ranked either by their association P-values from GWAS (from smallest to largest) or by their posterior probabilities of being associated with CD through incorporating rewiring information (from largest to smallest). By comparing the two rankings, we define the set of prioritized genes as those that are ranked higher in the latter one.

The performance of the prioritization approach was evaluated by the replication rate in the WTCCC GWAS. For the top-k list of genes with the smallest P-values in the NIDDK GWAS, we evaluated the replication rate of all k genes, the prioritized genes, and those genes not prioritized. We define a gene as 'replicable' if its association P-value in the WTCCC GWAS (validation cohort) was <0.05. The baseline replication rate is the proportion of replicable genes among the top-k genes. The replication rate of the prioritized genes is defined as the number of prioritized genes that are replicable, divided by the number of prioritized genes. The replication rate of not prioritized genes is similarly defined. The replication rate was calculated from the top 50 to 500 genes (with a step size of 50, see Fig. 3A). We can see that there is an increase of replication rate in the prioritized set of genes. It is worth noting that the prioritized genes at a moderate cut-off can achieve similar replication rate as genes at a more stringent cut-off without prioritization, so that genes with moderate P-values can be recovered by our prioritization method, without sacrificing the replication rate.

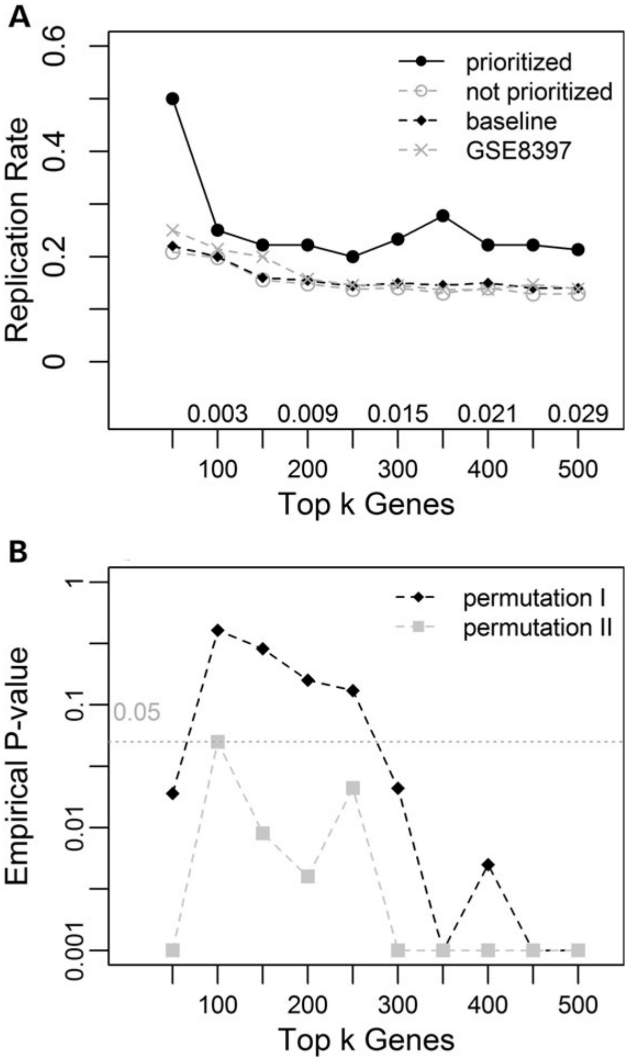

Figure 3. Replication rates between independent cohorts in Crohn's disease study. A gene is called replicable if its association P-value in the replication cohort is <0.05. (**A**). Black circle: prioritized genes. Grey circle: non-prioritized genes. Black diamond: baseline replication rates for the top k genes. Grey cross: prioritized genes with microarray data set GSE8397 (a microarray data set of Parkinson's disease). The numbers on the top of x-axis are the association P-value of the kth genes in the discovery cohort. (**B**). Empirical P-values of the replication rate of the prioritized gene set, derived from permutations. The dotted line indicates significance level at 0.05. The y-axis is drawn at the log_{10} scale but labelled with original P-values for ease of reading.

Besides a higher replication rate in the WTCCC cohort, some of the prioritized genes are mapped in CD susceptible loci that have been documented as meeting genome-wide significance level in the NHGRI GWAS Catalogue (25). More specifically, DCLRE1B and TMEM17 were reported to be associated with CD in a recent meta-analysis of inflammatory bowel disease with $>75\,000$ cases and controls (3); SOCS1 was shown to be associated with inflammatory bowel disease in the same study (3), and also reported in a large-scale meta-analysis of CD and psoriasis (26); SMAD3 and MAMSTR were identified in a meta-analysis of CD with 6333 cases and 15 056 controls (2); RPL7 and CPAMD8 were reported in a GWAS study of CD among individuals of Ashkenazi Jewish descent, with 907 cases and 2345 controls in the discovery stage, and 971 cases and 2124 controls in the replication study (27). Interestingly, none of the above mentioned genes had significant differential expressions in the microarray study (28).

These results demonstrate that our Markov random field framework of incorporating rewiring information can help better identify more replicable association signals. We now address the significance of the improvement in the following sections. Theoretical justification of the improvement is not a trivial task without specifying the underlying models of susceptible loci for complex disease. Instead, we constructed null distributions by randomizing rewiring network (29), and contrasted our results with two permutation settings and a third negative control by considering networks constructed from unrelated microarray data.

In permutation setting I, we constructed a null distribution of the rewiring network by permuting the case–control labels of microarray samples and recalculating the rewiring degree (see Materials and Methods). We did the permutations 1000 times. For each permuted network, we performed the same analysis as in the real data, including calculating the proportion of rewiring edges of $(+1, +1)$, $(+1, -1)$ and $(-1, -1)$, enrichment of rewiring edges in Reactome pathways, and prioritization of association signals, as detailed below.

First, we calculated the proportions of $(+1, +1)$ and $(+1, -1)$ edges in the rewiring network, and the significance of enrichment of each type of rewiring edge was evaluated by the empirical *P*-value, defined as the number of permutations with the corresponding proportion greater than that in the real data (Supplementary Material, Table S4). The results show that the excess of associated genes in rewiring edges in real data was statistically significant compared with the 1000 networks generated in permutation I (empirical *P*-values ranging from 0.013 to 0.022 at varying network thresholds).

Second, in each permutation, we calculated the fold change of rewiring edges in the 'immune system' as well as other pathways documented in the Reactome database. The Reactome pathway annotation has a hierarchical structure. To avoid redundancy in pathway annotations, only the first layer of pathway nodes were used, which has 17 general pathways (see Supplementary Material, Table S5 for a full list of the 17 pathways). We calculated the rewiring density of subnetwork defined by each pathway, and the corresponding fold change of density as opposed to the overall network. The statistical significance of the fold change was empirically evaluated by permutation I (see Supplementary Material, Table S5). Two pathways, 'immune system' and 'membrane trafficking', were nominally enriched with rewiring. The enrichment of rewiring edges in 'immune system' demonstrates that network rewiring is abundant in disease-related pathways.

Third, we used the permuted rewiring network to prioritize the association results of the NIDDK GWAS data and evaluated the replication rate in the WTCCC GWAS data accordingly (see Fig. 3B). When *k* was relatively small, although the replication rate was increased compared with the baseline level, the improvement was not significant. The empirical *P*-value became more significant as *k* increased. The permutation results demonstrated that for genes with moderate association evidence, the subset of genes prioritized by our method are indeed enriched with replicable signals.

It is possible that the increase in replication rate we observed might be attributed to differential expression of genes instead of network rewiring. We performed a second type of permutations to address this concern. In this setting, the disease samples and the control samples were permuted separately (see Materials and Methods), and as a result, the network structure was changed while differential expression remained the same as the real data. We did the permutation 1000 times, where the rewiring network was calculated and used for prioritization in each permutation. By contrasting the replication rate from the real data to the permutation results (see Fig. 3B), we can see that the permutation results were inferior to those from the real data even though differential expression was kept intact between the permutated and real data. This demonstrates that the improvement of replication rate was largely attributed to the appropriate modelling and incorporation of rewiring information, not to the overlapping information between rewiring and differential expression.

It is interesting to investigate what types of genes are rewired but not differentially expressed between disease and control conditions, and how they are related to disease aetiology. One possibility is that the protein activities of these genes are regulated by post-translational modifications, which may change in disease condition. For example, reduced phosphorylation level of SMAD3, which was prioritized by our method, was observed in mucosal samples from CD (30), while phosphorylation of SMAD3 is important for TGFB1 induced anti-inflammatory activities (31). Many in the prioritized gene set have documented post-translational modifications, although their relatedness with CD aetiology is not yet established.

Application to PD

Although our method was motivated and validated using data from CD, the computational framework can be applied to other diseases. To demonstrate its more general applicability, we applied the method to non-immune phenotypes. In this paper, we considered PD because of the availability of two GWAS data sets and one relevant microarray data set. The first PD GWAS data set (dbGaP study accession: phs000126.v1.p1) is an NIH funded genetic study aiming to identify risk loci for PD, which had a cohort of 857 cases and 867 controls. The second cohort (dbGaP study accession: phs000089.v3.p2) had 1713 PD cases and 3978 controls with Caucasian ancestry (32). We used the first one as discovery cohort and the second cohort (of larger sample size) as replication. We used a microarray data set (GEO accession: GSE8397) consisting of 47 brain tissue samples of PD cases and controls (33) to construct the PD rewiring network.

The PD GWAS signals were prioritized and the replication rate was evaluated in the same way as we did for CD (see Fig. 4). Although the replication rate of PD was lower than that in CD, there was indeed an increase of replication rate in the prioritized genes.

Moreover, some of the prioritized genes are mapped in loci with genome-wide significant association with PD (25). The association of MMRN1 was reported in a two-stage GWAS study of PD, with 1713 cases and 3978 controls in the discovery stage, and 3361 cases and 4573 controls in the replication stage (32). The loci were also reported in a recent meta-analysis of PD,

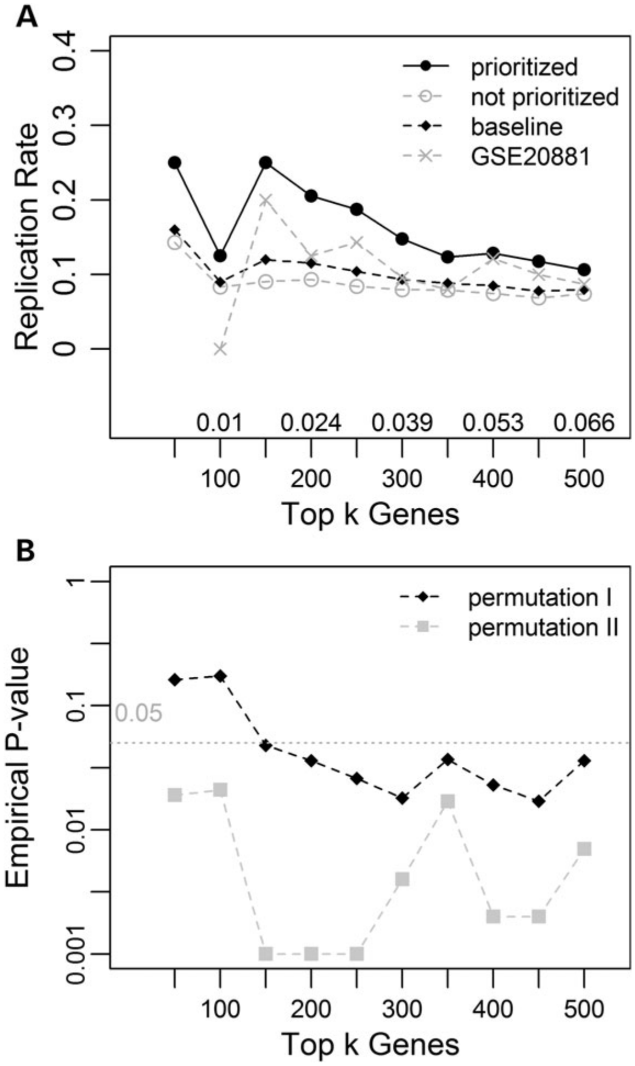

Figure 4. Replication rates between independent cohorts in Parkinson's disease study. A gene is called replicable if its association P-value in the replication cohort is <0.05. (**A**) Black circles: prioritized genes. Grey circles: non-prioritized genes. Black diamonds: baseline replication rates for the top k genes. Grey crosses: prioritized genes with microarray data set GSE20881 (a microarray data set of Crohn's disease). The numbers on the top of x-axis are the association P-value of the kth genes in the discovery cohort. (**B**) Empirical P-values of the replication rate of the prioritized gene set derived from permutations. The dotted line indicates significance level at 0.05. The y-axis is drawn at the log_{10} scale but labelled with original P-values for ease of reading.

with up to 16 452 cases and 48 810 controls (34). RIT2 and ZNF646 were identified in a GWAS study of PD with 4238 cases and 4239 controls in the discovery cohort, and 3738 cases and 2111 controls in the replication cohort (35). SIX1B was recently reported in a two-stage meta-analysis of PD, with 12 386 cases and 21 026 controls combining the two stages (36). These genes were not found to be differentially expressed in the microarray study (33).

Using the first type of permutations discussed above, the improvement of replication rate was not significant for the top 150 GWAS signals, but the improvement became significant as k increased. On the other hand, the statistical significance of the improvement based on the second type of permutations described above was significant for different k values. The comparable performance in gene prioritizations in PD demonstrates the generality of our method in prioritizing GWAS signals for non-immune-related phenotypes.

Disease-specific expression data are valuable for prioritization

Having demonstrated that incorporating rewiring information inferred from a case–control microarray data set of disease-related tissue could improve replication rate in both CD and PD GWAS studies, we now discuss the importance of the disease-specific information for prioritization. That is, whether we can still achieve improved replication rates if the network information is derived from an expression study unlikely related to disease of interest.

To investigate this question, we flipped the CD and PD microarray data sets, and applied the Markov random field model in prioritization. That is, we used the CD microarray data set to prioritize the GWAS signal of PD, and used the PD microarray data set to prioritize the GWAS signal of CD. The disease-related tissues are different (intestine for CD and brain for PD), thus the CD microarray data set would not be informative for association of PD, and vice versa. We would not expect that our method would lead to more replicable results, which was what we observed in the CD data (Fig. 3A). In the PD application, the replication rate was improved at several points ($k = 150$, 250 and 400) by incorporating the CD microarray data set (Fig. 4A), although none of those genes were replicable when a more stringent replication criterion was applied (P-value <0.01 in the replication cohort, see Supplementary Material, Fig. S3). Conversely, the improvement of the replication rate with disease-specific microarray data sets and its significance was retained for both CD and PD (see Supplementary Material, Figs S2 and S3). These results underscore the fact that the increase of replication rate was likely attributed to the rewiring information in the disease-specific rewiring network, instead of the common network connections in the human genome.

Functional gene set enrichment analysis of the prioritized genes

GWAS studies have revealed many disease-related pathways, especially in CD (37,38). Here, our purpose is to test whether the prioritized genes can be more enriched in disease-related pathways compared with the genes that are selected by GWAS signals alone. Thus, we used the hyper-geometric test for enrichment analysis (see Materials and Methods), where the test set is the prioritized genes in a disease, and the background set consists of genes with significant GWAS signals (P-value <0.05). We tested the gene set enrichment of the prioritized genes of CD and PD. The pathway database used is Reactome and GO biological process.

In the CD study, there were 780 genes with association P-values <0.05 (the contrast set) in the NIDDK cohort, while 124 of them were prioritized (test set, see Supplementary Material, Table S6 for the list of prioritized genes). Gene sets with P-values <0.05 are reported (see Supplementary Material, Table S8). Among the enriched gene sets, 'membrane lipid metabolic process' and 'phospholipid metabolism', part of lipid metabolism, were previously found to be enriched in GWAS signals in CD by Ballard *et al* (37) and Torkamani, *et al* (38). The enrichment of the 'Beta defensins' pathway appears to be new and not reported elsewhere. Beta defensins can promote adaptive immune response to micro-organisms by interaction with CCR6 (39). Meta-analysis of CD GWAS have established the association of CCR6 with CD at genome-wide significance level (3), but the underlying biological mechanism is yet unknown. Our results might be helpful in placing CCR6 in CD aetiology.

For PD, there were 379 genes with association *P*-values <0.05 (background set) and 74 were prioritized (test set, see Supplementary Material, Table S7). The prioritized genes were enriched in 'neurological system process', which is expected since PD is a neurological disorder. Nonetheless, the results can be helpful to pinpoint candidate genes involved in PD aetiology in this process (see Supplementary Material, Table S9). In addition, this pathway is not enriched for GWAS signals alone, where the test set is the genes with association *P*-values <0.05, and the background set includes all the genes in the study. However, none of the above gene sets could reach statistical significance after multiple test correction.

DISCUSSION

In this paper, we proposed a Markov random field approach to prioritizing GWAS signals by incorporating network rewiring information. Real-data analyses in CD and PD demonstrate that our method improves the replication rate between independent studies. Gene set enrichment analysis showed that the proposed method identified known pathways associated with these diseases, and also implicated potential pathways that have not been suggested by existing methods.

Our method differs from other network-based prioritization approachs by principle. While most other methods are based on the 'guilt by association' assumption, we propose a 'guilt by rewiring' approach, which prioritizes genes with network rewiring by comparing the co-expression networks in disease samples and control samples. There are several advantages of the proposed approach. First, the network information incorporated here is inferred from microarray experiments, which are less biased and more comprehensive than protein interaction networks and pathway knowledge databases. The bias in the latter data sources might weaken the power of GWAS in uncovering novel loci with disease association, since they put more weight on well-studied genes. In the network rewiring analysis, each gene is represented as long as expression information is available. Second, the prioritized genes in our method have dynamic behaviour in disease state, and this is valuable because they may be potential biomarkers or drug targets. Third, the prioritization results may be helpful in designing replication, so that selected genes or loci with moderate GWAS signals can be included in follow-up studies, while not sacrificing the replication rate. The software, an R package named 'GBR', applying our method is available at http://bioinformatics.med.yale.edu/group/.

Despite the advantages of our guilt by rewiring approach, the prioritized genes bear an inherent limitation: possible inflations due to physical proximity of genes in the genome. For example, the enrichment of biological pathways might be a consequence of proximal genes annotated in the same pathway instead of the prioritization scheme. This is inherited from the input GWAS signals, which might be inflated by gene proximity as well. However, in the application to CD and PD, the prioritized genes in the enriched pathways are scattered in different chromosome arms (see Supplementary Material, Tables S8 and S9), thus excluding inflation as a confounding factor here. Nonetheless, the interpretation of the prioritized genes should be made with awareness of the potential problem in future application of this method.

To study the interplay between genotype and network rewiring pattern, the best approach would be to investigate samples where both GWAS and expression data are available. However, in the current study, the microarray data and GWAS data were from different samples. More insight can be gained when matched data are available. In the future, we would like to extend the model to incorporate other network information such as eQTL and transcription factor binding, to uncover the relationship between the rewiring diagram of transcriptional regulatory network and complex disease.

MATERIALS AND METHODS

Data sets of Genome Wide Association Studies

The GWAS data sets of CD and PD are described in Supplementary Material, Table S1, including sample size, genotyping platform and other information.

Gene assignment

To determine disease association at the gene level, we used two methods based on data availability level (see Supplementary Material, Table S1 for detail). Principal component analysis (PCA) was applied when the genotype data were available (40). Among all markers that are within 10 kb upstream or downstream of the gene start/end region, we first performed PCA, and then used the top *l* principal components ($PC_1, PC_2, \ldots, PC_l$) that explained at least 95% of the total variance for logistic regression [see Eq. (1)]. Besides, a model with only population structure components was fitted [Eq. (2)], and the *P*-value of the gene was determined by the likelihood ratio test between these two models.

$$log\frac{p}{1-p} = a_0 + \sum_{i=1}^{l} a_i \mathrm{PC}_i + \sum_{i=1}^{4} b_i \mathrm{PS}_i. \tag{1}$$

$$log\frac{p}{1-p} = a_0 + \sum_{i=1}^{4} b_i \mathrm{PS}_i. \tag{2}$$

For data sets where individual genotype data were not available, the minSNP method was employed (40). Suppose *n* SNPs are assigned to the gene, with association *P*-values, $p_1, p_2, \ldots, p_n$, and $p_{(1)}$ is the minimum. In the minSNP method, the gene level association *P*-value is

$$p = 1 - (1 - p_{(1)})^n. \tag{3}$$

Gene expression

The gene expression data set was downloaded from GEO (GSE20881, GSE8397 see Supplementary Material, Table S2 for details). In the CD data set (GSE20881), the gene expression levels of 172 biopsies from 53 CD patients and 31 control samples were measured on the Agilent Whole Human Genome Microarray (28). There were originally 44 290 probe sets on the Agilent platform, and we pre-processed the data in the following steps: one sample was excluded because the corresponding expression profile differed substantially from the others (Supplementary Material, Fig. S1); for genes that were mapped by multiple probe sets, we excluded the gene if the Pearson correlation coefficient between the multiple probe sets was <0.27, which corresponds to the 0.99 quantile of Student t-distribution [Eq. (4), n_m is the number of microarray samples]. Otherwise, one single probe set was selected to represent the gene (see Supplementary Material, Method S1 for details). In total, 15 041 genes remained in our analysis, and 12 007 of them overlapped with those in the CD GWAS studies.

$$T = r\sqrt{\frac{n_m - 2}{1 - r^2}} \sim t_{n_m - 2}. \qquad (4)$$

The PD microarray data set (GSE8397) was processed similarly with 11 106 genes remaining after quality control, and the number of genes overlapping with the GWAS studies was 8987.

Measurement of network rewiring by differential co-expression

The PCC was calculated for each pair of genes in CD samples and control samples, separately. Let r_{ij}^{CD} denote the PCC of genes i and j in the CD samples, and $r_{ij}^{\mathrm{control}}$ that in the control samples. Previously, the difference between PCCs was used to measure differential rewiring (20). Hu *et al.* (41) showed, by simulation, applying Fisher transformation [Eq. (5)] improved the power to identify differentially rewired genes. Here, we used Fisher's test of difference between two correlation coefficients, which considers both the change of PCC level and effect of sample size [Eq. (6)]. The test statistic approximately follows standard normal distribution under the null hypothesis of no difference between the PCC levels between patients and controls. Thus, the rewiring information, rewire_{ij}, is defined as a value between 0 and 1, with larger value indicating more dramatic rewiring effect.

$$F(r) = \frac{1}{2} ln \frac{1 + r}{1 - r}. \qquad (5)$$

$$\mathrm{rewire}_{ij} = P(|X| \le |\frac{F(r^{\mathrm{CD}}) - F(r^{\mathrm{control}})}{\sqrt{\frac{1}{n_{\mathrm{CD}} - 3} + \frac{1}{n_{\mathrm{control}} - 3}}}|), X \sim N(0, 1). \qquad (6)$$

Network dichotomization

The rewiring and co-expression networks are both weighted networks, with weights ranging between 0 and 1. However, the weights, rewiring information and absolute PCC value, are distinct concepts and not comparable by nature. To facilitate the comparison, we dichotomized the two networks in such a way that they had the same network density. In detail, the rewiring information of all gene pairs was ranked, and the 0.9, 0.95 and 0.99 quantiles were chosen as the hard threshold. The resultant network densities were 0.1, 0.05 and 0.01, respectively. The static co-expression network is dichotomized likewise by hard-thresholding on the absolute PCC values.

Markov random field modelling

To prioritize disease-associated genes with network rewiring, we utilized an Hidden Markov random field model to formulate the problem. In the network, each node is a gene, with an association label ω_i, either $+1$ (associated) or -1 (not associated). A network configuration is the label vector of all nodes in the network, $(\omega_1, \omega_2, \ldots, \omega_N)$, where N is the number of genes considered. Two genes are connected ($e_{ij} = 1$) if they were co-expressed either in the disease state or healthy state. The threshold used to dichotomize the co-expression network was chosen by power law distribution (see Supplementary Material, Methods). The degree of rewiring (rewire_{ij}) is described in the previous section. The distribution of network configuration is defined as follows:

$$\begin{aligned} P(\omega_1, \omega_2, \cdots, \omega_N) = \frac{1}{Z} exp(-h \sum_{i=1}^{N} I(\omega_i = 1) \\ + \tau_1 \sum_{e_{ij}=1} \mathrm{rewire}_{ij} \cdot I(\omega_i = 1, \omega_j = 1) \\ - \tau_2 \sum_{e_{ij}=1, \mathrm{rewire}_{ij} > \delta} \mathrm{rewire}_{ij} \cdot I(\omega_i = -1, \omega_j = -1), \qquad (7) \end{aligned}$$

where (h, τ_1, τ_2) are hyper-parameters, $I(\cdot)$ is an indicator function and Z is the partition function. In Eq. (7), the $\tau.$ is weighting parameters with positive values, which determine the influences of different types of edges. The impacts of these parameters can be better explained with Eq. (8), which shows the conditional probability of association state of gene i. This probability is composed of three parts: (i) h is a constant defining the probability of being disease associated if the gene is isolated thus no network information can be incorporated; (ii) τ_1 indicates the contribution of rewiring degree of 'associated' neighbours; while (3) τ_2 indicates those of 'non-associated' neighbours. Suppose the association states of the neighbours of gene i are fixed, the larger τ_1 and τ_2 are, the more likely gene i is disease associated.

$$\begin{aligned} log \frac{p(\omega_i = 1|\omega_{-i})}{p(\omega_i = -1|\omega_{-i})} = -h + \tau_1 \sum_{e_{ij}=1, \omega_j=1} \mathrm{rewire}_{ij} \\ + \tau_2 \sum_{e_{ij}=1, \omega_j=-1, \mathrm{rewire}_{ij} > \delta} \mathrm{rewire}_{ij}. \qquad (8) \end{aligned}$$

Here ω_{-i} stands for all genes in the network except gene i. In our analysis, δ was set to 0.95, and a rewiring degree less than that indicates that the difference of PCC between patients and control samples is not significant. The odds of a gene to be associated with disease will increase with larger rewiring degree with its neighbours. The underlying biological assumption is that a group of erroneous gene interactions, which are present in one condition and absent in the other, probably reflect organizational changes of the cellular networks under different disease conditions.

Given the network structure and the association signals, the posterior probability of the network configuration can be inferred through Bayesian framework:

$$P(\Omega|Y) \propto P(Y|\Omega)P(\Omega). \qquad (9)$$

The observed data $Y = (y_1, y_2, ..., y_N)$ are the normalized scores corresponding to the P-values in GWAS studies: $y_i = \Phi^{-1}(1 - p_i)$, where Φ is the cumulative distribution function of a standard normal variable. Under the null hypothesis that the gene is not associated with the disease, its P-value follows a *Uniform*(0,1) distribution. Thus, $P(y_i|\omega_i = -1) \sim N(0, 1)$. Under the alternative hypothesis, i.e. the association state is '+1', we follow Chen *et al.* by assuming $P(y_i|\omega_i = 1) \sim N(\mu_i, \sigma_i^2)$, and assign Φ conjugate priors for μ_i and σ_i^2 [Eq. (10)].

$$\mu_i|\sigma_{i^2} \sim N(\bar{\mu}, \sigma_i^2/a), \sigma_i^2 \sim \text{InverseGamma}(\upsilon/2, \upsilon d/2). \qquad (10)$$

The hidden states can be inferred by the iterated conditional modes algorithm (42).

Although this modelling framework is similar to that by Chen *et al.* (23), the proposed approach is different in principle. First, the previous approach assumes connected genes in a pathway tend to share association states, which is a 'guilt by association' approach in the general sense. Second, the current approach incorporates the network structure at a systems level, and not restricted to connections defined within annotated pathways as proposed in the previous approach. Indeed, dynamic changes tend to involve genes between pathways, rather than within known pathways.

Choice of hyper-parameters

There are two sets of hyper-parameters in our model, including network parameters (h, τ_1, τ_2) and GWAS parameters ($\bar{\mu}, a, \upsilon, d$), respectively. The parameters (τ_1, τ_2) reflect the context-dependent contribution of network rewiring to the configuration distribution. The change of energy function caused by assigning node i to '+1' from '−1' [Eq. (11)] can be easily derived from [Eq. (7)]. We fixed both τ_1 and τ_2 as 1, since we assume the rewiring with both associated and non-associated genes increase the probability that this gene is associated.

$$\tau_1 \sum_{\omega_j=1, e_{ij}=1} \text{rewire}_{ij} + \tau_2 \sum_{\omega_j=-1, e_{ij}=1, \text{rewire}_{ij}>\delta} \text{rewire}_{ij} - h. \qquad (11)$$

The parameter h, a negative value, determines the distribution of network configuration when neither GWAS nor gene expression data are available. When (τ_1, τ_2) are fixed, a larger value of h favours network configurations with more nodes labelled as 'not associated'. We choose h empirically (see details in Supplementary Material, Method). The GWAS parameters have been previously discussed (23), where the authors noted the results are not sensitive to these parameters based on simulation studies. In this article, we adopted the same set-up (1,3,10).

Permutations

In permutation I, we reshuffled the case–control labels of the microarray samples. In permutation II, the case and control samples were shuffled separately. In the microarray data set of control samples, the expression profile of gene i is $\{x_{i1}, x_{i2}, \ldots, x_{is}\}$. Let $\{r_{i1}, r_{i2}, \ldots, r_{is}\}$ denote a permutation of $\{1, 2, \ldots, s\}$, the expression profile of gene i in control samples was permuted to $\{x_{ir_{i1}}, x_{ir_{i2}}, \cdots, x_{ir_{is}}\}$. The expression profiles of disease samples were shuffled in the same way. In the real data, x_{ik} and x_{jk} are expression levels of gene i and gene j from the kth patient, but the pairing is broken in the permutation due to the randomization process. In this way, the rewiring pattern is shuffled while the differential expression is preserved for each gene. In both permutation settings I and II, 1000 networks were generated based on the permuted data.

Gene set enrichment analysis

Suppose that there are M genes in the background set, and m of those genes are prioritized. The number of overlap genes of the background set and the prioritized set with a functional gene set is M_p and m_p, respectively. In the hyper-geometric test, the enrichment P-value was calculated as follows:

$$P_{\text{HG}} = \frac{C_{M_p}^{m_p} C_{M-M_p}^{m-m_p}}{C_M^m}. \qquad (12)$$

For permutations, we randomly sampled m genes from the background set for 1000 times, and calculated the intersection of the random set and the functional gene set. Empirical P-value was defined as $\frac{\sum_{i=1}^{1000} I(m_i \geq m_p)}{1000}$, where m_i is the number of overlapping genes in the random set.

Binomial test

The significance of enrichment of $(+1, +1)$ and $(+1, -1)$ edges in the rewiring and the static network was addressed by the binomial test. Supposed the expected proportion of $(+1, +1)$ is θ, the number of all edges and $(+1, +1)$ edges are N_T and N_{11}, respectively, the significance of enrichment is $\sum_{k \geq N_{11}} C_{N_T}^k \theta^k (1 - \theta)^{N_T - k}$. The P-values were reported as 0 when they fell <2.220446e-16, which is the smallest floating-point number in R environment.

Software

The software of the guilt by rewiring approach is available at http://bioinformatics.med.yale.edu/group/.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank the International IBD Genetics Consortium (IIBDGC) for providing the GWAS data.

Conflict of Interest statement. None declared.

FUNDING

This work was supported in part by the National Institutes of Health (R01-GM59507, UL1 RR024139, and P01CA154295 to L.H. and H.Z., U01-DK062422, R01-DK092235, KL2RR024138, and U01-DK62429 to J.C., 1K25AR063761 to M.C.); the Department of Veterans Affairs (VA Cooperative Studies Program to L.H. and H.Z.), the National Science Foundation (DMS 1106738 to H.Z.); the Cancer Prevention and Research Institute of Texas (RP101251 to M.C.); the National Aeronautics and Space Administration (NNJ05HD36G to M.C.) and the Clinical and Translational Science Award (UL1 RR024139).

REFERENCES

1. Manolio, T.A., Brooks, L.D. and Collins, F.S. (2008) A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.*, **118**, 1590–1605.
2. Franke, A., McGovern, D.P.B., 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.
3. Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Philip Schumm, L., Sharma, Y., Anderson, C.A. *et al.* (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, **491**, 119–124.
4. Goldstein, D.B. (2009) Common genetic variation and human traits. *N. Engl. J. Med.*, **360**, 1696–1698.
5. Hardy, J. and Singleton, A. (2009) Genomewide Association Studies and human disease. *N. Engl. J. Med.*, **360**, 1759–1768.
6. Chen, J., Bardes, E.E., Aronow, B.J. and Jegga, A.G. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucl. Acids Res.*, **37**, W305–W311.
7. Saccone, S.F., Saccone, N.L., Swan, G.E., Madden, P.A.F., Goate, A.M., Rice, J.P. and Bierut, L.J. (2008) Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. *Bioinformatics*, **24**, 1805–1811.
8. Sun, J., Jia, P., Fanous, A.H., Webb, B.T., van den Oord, E.J.C.G., Chen, X., Bukszar, J., Kendler, K.S. and Zhao, Z. (2009) A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases–schizophrenia as a case. *Bioinformatics*, **25**, 2595–2602.
9. Franke, L., Bakel, H.v., Fokkens, L., de Jong, E.D., Egmont-Petersen, M. and Wijmenga, C. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.
10. Lee, I., Blom, U.M., Wang, P.I., Shim, J.E. and Marcotte, E.M. (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.*, **21**, 1109–1121.
11. Moreau, Y. and Tranchevent, L.-C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.*, **13**, 523–536.
12. Oti, M. and Brunner, H.G. (2007) The modular nature of genetic diseases. *Clin. Genet.*, **71**, 1–11.
13. Wu, X., Jiang, R., Zhang, M.Q. and Li, S. (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol.*, **4**, 189.
14. Yang, P., Li, X., Wu, M., Kwoh, C.-K. and Ng, S.-K. (2011) Inferring gene-phenotype associations via global protein complex network propagation. *PLoS ONE*, **6**, e21502.
15. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. and Sharan, R. (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.*, **6**, e1000641.
16. Luscombe, N.M., Madan Babu, M., Yu, H., Snyder, M., Teichmann, S.A. and Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.
17. Ideker, T. and Krogan, N.J. (2012) Differential network biology. *Mol. Syst. Biol.*, **8**, 565.
18. Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.-K., Chuang, R., Jaehnig, E.J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M. *et al.* (2010) Rewiring of genetic networks in response to DNA damage. *Science*, **330**, 1385–1389.
19. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucl. Acids Res.*, **39**, D1005–D1010.
20. Hudson, N.J., Reverter, A. and Dalrymple, B.P. (2009) A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Comput. Biol.*, **5**, e1000382.
21. Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q. and Wrana, J.L. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.*, **27**, 199–204.
22. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucl. Acids Res.*, **37**, D619–D622.
23. Chen, M., Cho, J. and Zhao, H. (2011) Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLoS Genet.*, **7**, e1001353.
24. Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M.M., Datta, L.W. *et al.* (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.*, **39**, 596–604.
25. Hindorff, L., Sethupathy, P., Junkins, H., Ramos, E., Mehta, J., Collins, F. and Manolio, T. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
26. Ellinghaus, D., Ellinghaus, E., Nair, R.P., Stuart, P.E., Esko, T., Metspalu, A., Debrus, S., Raelson, J.V., Tejasvi, T., Belouchi, M. *et al.* (2012) Combined analysis of genome-wide association studies for Crohn disease and psoriasis identifies seven shared susceptibility loci. *Am. J. Hum. Genet.*, **90**, 636–647.
27. Kenny, E.E., Pe'er, I., Karban, A., Ozelius, L., Mitchell, A.A., Ng, S.M., Erazo, M., Ostrer, H., Abraham, C., Abreu, M.T. *et al.* (2012) A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. *PLoS Genet.*, **8**, e1002559.
28. Noble, C.L., Abbas, A.R., Lees, C.W., Cornelius, J., Toy, K., Modrusan, Z., Clark, H.F., Arnott, I.D., Penman, I.D., Satsangi, J. *et al.* (2010) Characterization of intestinal gene expression profiles in Crohn's disease by genome-wide microarray analysis. *Inflamm. Bowel Dis.*, **16**, 1717–1728.
29. Rossin, E.J., Lage, K., Raychaudhuri, S., Xavier, R.J., Tatar, D., Benita, Y., Cotsapas, C. and Daly, M.J. and International Inflammatory Bowel Disease Genetics, C. (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.*, **7**, e1001273.
30. Monteleone, G., Kumberova, A., Croft, N.M., McKenzie, C., Steer, H.W. and MacDonald, T.T. (2001) Blocking Smad7 restores TGF-β1 signaling in chronic inflammatory bowel disease. *J. Clin. Invest.*, **108**, 601–609.
31. Yang, X., Letterio, J.J., Lechleider, R.J., Chen, L., Hayman, R., Gu, H., Roberts, A.B. and Deng, C. (1999) Targeted disruption of SMAD3 results in impaired mucosal immunity and diminished T cell responsiveness to TGF-beta. *EMBO J.*, **18**, 1280–1291.
32. Simon-Sanchez, J., Schulte, C., Bras, J.M., Sharma, M., Gibbs, J.R., Berg, D., Paisan-Ruiz, C., Lichtner, P., Scholz, S.W., Hernandez, D.G. *et al.* (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.*, **41**, 1308–1312.
33. Moran, L.B., Duke, D.C., Deprez, M., Dexter, D.T., Pearce, R.K.B. and Graeber, M.B. (2006) Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. *Neurogenetics*, **7**, 1–11.

34. Lill, C., Roehr, J., McQueen, M., Kavvoura, F., Bagade, S., Schjeide, B., Schjeide, L., Meissner, E., Zauft, U., Allen, N. *et al.* (2012) Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: the PDGENE database. *PLoS Genet.*, **8**, e1002548.
35. Pankratz, N., Beecham, G., DeStefano, A., Dawson, T., Doheny, K., Factor, S., Hamza, T., Hung, A., Hyman, B., Ivinson, A. *et al.* (2012) Meta-analysis of Parkinson's disease: identification of a novel locus, RIT2. *Ann. Neurol.*, **71**, 370–384.
36. International Parkinson's Disease Genomics Consortium (IPDGC), Welcome Trust Case Control Consortium 2 (WTCCC2). (2011) A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS Genet.*, **7**, e1002142.
37. Ballard, D., Abraham, C., Cho, J. and Zhao, H. (2010) Pathway analysis comparison using Crohn's disease Genome Wide Association Studies. *BMC Med. Genomics*, **3**, 25.
38. Torkamani, A., Topol, E. and Schork, N. (2008) Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, **92**, 265–272.
39. Yang, D., Chertov, O., Bykovskaia, S., Chen, Q., Buffo, M., Shogan, J., Anderson, M., Schröder, J., Wang, J. and Howard, O. (1999) Beta-defensins: linking innate and adaptive immunity through dendritic and T Cell CCR6. *Science*, **286**, 525–528.
40. Ballard, D.H., Cho, J. and Zhao, H. (2010) Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genet. Epidemiol.*, **34**, 201–212.
41. Hu, R., Qiu, X., Glazko, G., Klebanov, L. and Yakovlev, A. (2009) Detecting intergene correlation changes in microarray analysis: a new approach to gene selection. *BMC Bioinfor.*, **10**, 20.
42. Besag, J. (1986) On the statistical analysis of dirty pictures. *J. R. Statist. Soc. B*, **48**, 259–302.