Reconstruction of a Functional Human Gene Network, with an Application for Prioritizing Positional Candidate Genes

Lude Franke,[1] Harm van Bakel,[1] Like Fokkens,[1] Edwin D. de Jong,[2] Michael Egmont-Petersen,[3] and Cisca Wijmenga[1]

[1]Complex Genetics Section, Department of Biomedical Genetics–Department of Medical Genetics, University Medical Centre Utrecht, and [2]Large Distributed Databases Group, Institute of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands; and [3]Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

Most common genetic disorders have a complex inheritance and may result from variants in many genes, each contributing only weak effects to the disease. Pinpointing these disease genes within the myriad of susceptibility loci identified in linkage studies is difficult because these loci may contain hundreds of genes. However, in any disorder, most of the disease genes will be involved in only a few different molecular pathways. If we know something about the relationships between the genes, we can assess whether some genes (which may reside in different loci) functionally interact with each other, indicating a joint basis for the disease etiology. There are various repositories of information on pathway relationships. To consolidate this information, we developed a functional human gene network that integrates information on genes and the functional relationships between genes, based on data from the Kyoto Encyclopedia of Genes and Genomes, the Biomolecular Interaction Network Database, Reactome, the Human Protein Reference Database, the Gene Ontology database, predicted protein-protein interactions, human yeast two-hybrid interactions, and microarray coexpressions. We applied this network to interrelate positional candidate genes from different disease loci and then tested 96 heritable disorders for which the Online Mendelian Inheritance in Man database reported at least three disease genes. Artificial susceptibility loci, each containing 100 genes, were constructed around each disease gene, and we used the network to rank these genes on the basis of their functional interactions. By following up the top five genes per artificial locus, we were able to detect at least one known disease gene in 54% of the loci studied, representing a 2.8-fold increase over random selection. This suggests that our method can significantly reduce the cost and effort of pinpointing true disease genes in analyses of disorders for which numerous loci have been reported but for which most of the genes are unknown.

The completion of various genome-sequencing projects and large-scale genomic studies has led to a wealth of available biological data. It is anticipated that this information will revolutionize our insight into the molecular basis of most common diseases by making it easier and quicker to identify genes with variants that predispose to disease (i.e., disease genes). At the moment, we are faced with many disease susceptibility loci, resulting from linkage or cytogenetic analyses, that cover extensive genomic regions. Usually, when the genes in these loci are assessed, positional candidate genes become apparent that can be linked to the phenotype being studied on the basis of their biological function.

However, the most obvious functional candidate gene from a disease locus does not always prove to be involved in the disease.[e.g.,1–5] Often, genes that would not have been predicted to be disease causing prove to be the true disease gene—for example, the *BRCA1* gene in early-onset breast cancer.[6] Moreover, although these disease genes might have been assigned biological functions, it is not always evident how these functions relate to disease. Finally, genes with unknown functions are often overlooked, as attention is paid only to well-studied genes for which functions and interactions have been identified or implicated, some of which can be related to the disease pathogenesis. For example, in Fanconi anemia, at least 10 disease genes were identified,[7] but only a few had a known function. However, follow-up research[8–10] revealed that five of those genes function in the same protein complex. Another example is limb-girdle muscular dystrophy, in which many of the disease genes encode for proteins that are part of the dystrophin complex.[11] This emphasizes the importance of taking an unbiased approach to assessing positional candidate genes.

Faced with the absence of complete functional information for the majority of genes in susceptibility loci, it is difficult to prioritize the positional candidate genes correctly for further sequence or association analysis. However, high-throughput genomic work has now yielded relatively unbiased genomewide data sets[12–15] that comprise known metabolic, regulatory, functional,

Received December 9, 2005; accepted for publication March 14, 2006; electronically published April 25, 2006.

Address for correspondence and reprints: Dr. Cisca Wijmenga, Complex Genetics Section, DBG–Department of Medical Genetics, Stratenum 2.117, University Medical Centre Utrecht, PO Box 85060, 3508 AB Utrecht, The Netherlands. E-mail: t.n.wijmenga@med.uu.nl

Am. J. Hum. Genet. 2006;78:1011–1025. 0002-9297/2006/7806-0011$15.00

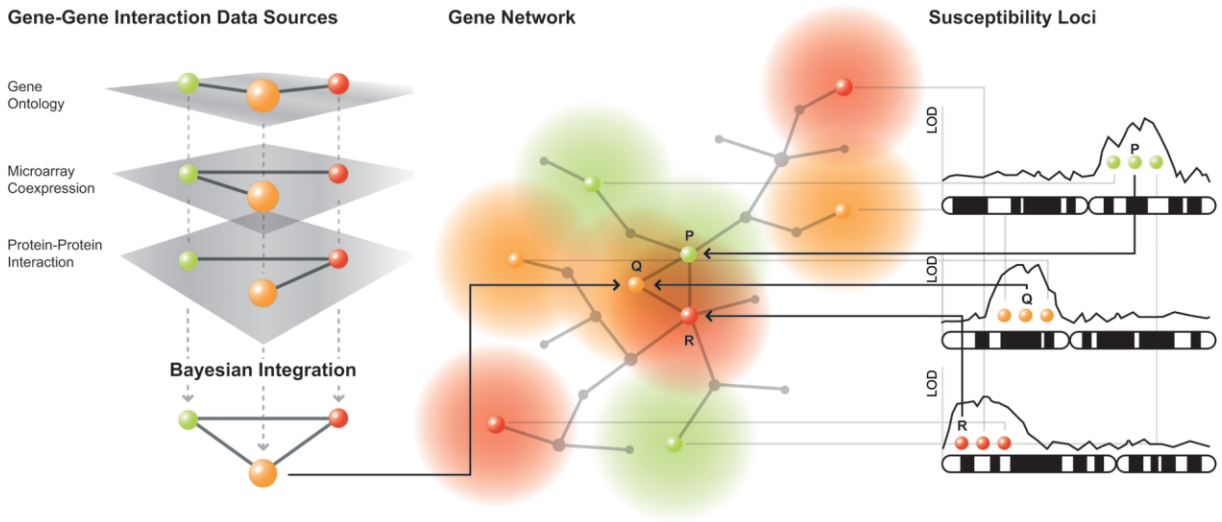

Figure 1 Basic principles of the prioritization method for positional candidate genes with the use of a functional human gene network. The method integrates different gene-gene interaction data sources in a Bayesian way (*left panel*). Subsequently, this gene network is used to prioritize positional candidate genes, with all genes assigned an initial score of zero. In the example (*right panel*), three different susceptibility loci are analyzed, each containing a disease gene (P, Q, or R) and two nondisease genes. In each locus, the three positional candidate genes increase the scores of nearby genes in the gene network, by use of a kernel function that models the relationship between gene-gene distance and score effect. Genes within each locus are ranked on the basis of their eventual effect score, corrected for differences in the topology of the network (see the "Material and Methods" section).

and physical interactions. There is, however, little integration of these diverse data sets into a coherent view of possible gene and protein interactions that can be used to investigate relationships between genes in different genetic loci. We have tried to address this problem by developing a functional human gene network that comprises known interactions derived from the Biomolecular Interaction Network Database (BIND),[12] the Human Protein Reference Database (HPRD),[13] Reactome,[15] and the Kyoto Encyclopedia of Genes and Genomes (KEGG).[14]

Since these data sets contain a limited number of known interactions, we implemented a Bayesian framework to complement these relationships with a large number of predicted interactions by relying on evidence for putative gene relationships based on biological process and molecular function annotations from the Gene Ontology database (GO).[16] We further incorporated experimental data—namely, coexpression data derived from ~450 microarray hybridizations from the Stanford Microarray Database (SMD)[17] and the NCBI Gene Expression Omnibus (GEO),[18] along with human yeast two-hybrid (Y2H) interactions[19] and interactions based on orthologous high-throughput protein-protein interactions from lower eukaryotes.[20]

Our interaction network was then used to test whether we could rank the best positional candidates in susceptibility loci on the basis of their interactions, assuming that the causative genes for any one disorder will be involved in only a few different biological pathways. This would be apparent in our network as a clustering of genes from different susceptibility loci, resulting in shorter gene-gene connections between disease genes than one would expect by chance (fig. 1). Our method (called "Prioritizer") analyzes susceptibility loci and investigates whether genes from different loci can be linked to each other directly[21] or indirectly.[22] When we constructed artificial loci of varying size around susceptibility loci from 96 different genetic disorders (each containing at least three loci) and used Prioritizer in our most comprehensive gene network to rank the positional candidate genes for each locus, we were able to significantly increase the chance of detecting disease genes.

Material and Methods

Functional Gene Network Reconstruction

As a basis for the gene network, we used annotations from Ensembl,[23] version 32.35, resulting in 20,334 known genes that physically map within the autosomes or chromosome X or Y. This yielded 206,725,611 potential gene-gene interactions.

On the basis of this set of genes, a comprehensive "gold standard" set of validated direct gene-gene relationships (true positives) was determined using both BIND (September 15, 2005) and HRPD (September 15, 2005) to extract human, curated protein-protein interactions, the proteins of which were mapped to Ensembl gene identifiers. In addition, all hu-

man pathways from Reactome (September 15, 2005) and KEGG (September 15, 2005) were used to derive direct interactions that were of transcriptional, physical, or metabolic origin, since pathways are usually composed of genes and proteins that interact with each other in various ways. We chose to allow interactions of physical, metabolic, and regulatory origin to be included within our network, because, for instance, mutations in either one of two genes encoding proteins in the same metabolic pathway or protein complex could lead to the same disease phenotype.

Because the true-positive gold standard only describes a limited number of relationships between a limited number of genes, we also used data from GO, coexpression data derived from microarray experiments, conserved protein-protein high-throughput data, and human Y2H interaction data to predict interactions of the remaining gene pairs. We used a Bayesian classifier, because these four types of data were of varying reliability and only contained information about a subset of the data. The classifier allows for combining dissimilar data sets, can deal with missing data, and uses conditional probabilities that can be well interpreted and that control for the varying reliability of the data sets.[24–29]

Training of Bayesian Classifier on Gold Standard

For the prediction of interactions, we used a Bayesian classifier type that assumed all data sets had been binned. This operation was performed for each gene pair, and it determined, for each data set, to which bin the pair belongs. Because the number of bins per data set was limited, each bin contained many gene pairs. Subsequently, for each bin, we determined the likelihood ratio between the proportion of gene pairs known to interact and the proportion of gene pairs known not to interact. This measure indicates whether there is an over- or an underrepresentation of truly interacting gene pairs in the bin, which specifies the conditional probability estimates of the Bayesian classifier; thus, training of the classifier is straightforward.

However, to be able to train the classifier by determining likelihood ratios of sets of gene pairs, it was crucial that the gold standard, containing the aforementioned well-defined set of curated true-positive gene pairs, be complemented with a set of gene pairs for which there is strong evidence that they, or the proteins they encode, do not functionally interact (true negatives). As has been discussed by others,[30] the construction of this true-negative reference set is problematic, because it is impossible to be certain that two genes (i.e., their protein products) do not interact. However, by assuming that genes encoding for proteins localized within different cellular compartments are, in general, unrelated, it is possible to make a list of gene pairs that are unlikely to interact. The GO Cellular Component annotations were used to yield groups of gene pairs that have exclusive cellular component annotations. To overcome a strong selection bias in the classifier toward well-annotated genes (details provided in appendix A [online only]), only the 5,105 genes that were part of a true-positive gene pair at least three times were allowed to form true-negative gene pairs. We chose combinations of cellular organelles that were highly underrepresented ($\chi^2 = 2{,}490$; $P < 10^{-10}$) within the true-positive set, which resulted in gene pairs for the following combinations: nucleus and extracellular matrix, protein complex and Golgi apparatus, protein complex and Golgi stack, non–membrane-bound organelle and Golgi stack, non–membrane-bound organelle and extracellular space, non–membrane-bound organelle and Golgi apparatus, extracellular region and organelle membrane, mitochondrion and extracellular matrix, extracellular space and organelle membrane, extracellular space and Golgi stack, organelle membrane and extracellular matrix, extracellular matrix and Golgi stack, extracellular matrix and ubiquitin ligase complex, and ubiquitin ligase complex and Golgi stack.

Preprocessing and Binning of Data Sets

To allow for Bayesian integration, the GO data, microarray coexpression data, and orthologous and human protein-protein interactions data were preprocessed and binned. Biological Process and Molecular Function GO annotations were derived from Ensembl, and two measures of relatedness for each of the two data sets were determined, resulting in a total of four different GO measures of relatedness. First, we determined, for each Biological Process GO term, how many of the genes had been assigned this term. Then, we determined which Biological Process GO terms were shared between the two components of each gene pair, for all the pairs. This led to the shared GO term that was annotated in the least number of genes, and its frequency of occurrence was used as a measure. GO terms GO:0000004 (biological process unknown) and GO:0005554 (molecular function unknown) were discarded, since genes that shared either of these highly unspecific terms should not be related to each other on the basis of this information. The same procedure was performed to generate the first measure of Molecular Function GO relatedness.

The second measure determined the maximal hierarchical depth at which a gene pair shared a Biological Process GO term. This hierarchical depth was defined as the shortest number of branches necessary to go from one Biological Process GO term back to the GO root. The same method was used to generate the maximum hierarchical depth of the Molecular Function GO sharing measure.

Coexpression between genes was determined in microarray data sets from GEO and SMD. Individual data sets comprised an experiment that contained at least 10 hybridizations. To ensure that the quality of the intensity measurements was reliable, various filtering steps were performed to exclude spots with low signal-to-noise ratios.[31] Within the SMD data sets, intensity spots were filtered out that were either missing or contaminated, and the mean intensity of spots had to be at least 2.5 times higher than the average background signal of the microarray. Since GEO contains both ratiometric and Affymetrix single-spot intensity microarray data sets, we used different filtering strategies. The 5% of genes with the lowest maximal intensity were removed from the Affymetrix data sets. For both SMD and GEO, expression ratios were $\log_2$ transformed. Microarray features missing $\geq 25\%$ of expression measurements in a data set after filtering were excluded. All features were assigned Ensembl gene identifiers by comparing their sequences to Ensembl transcripts with the use of SSAHA.[32]

To determine which gene pairs showed coexpression, the

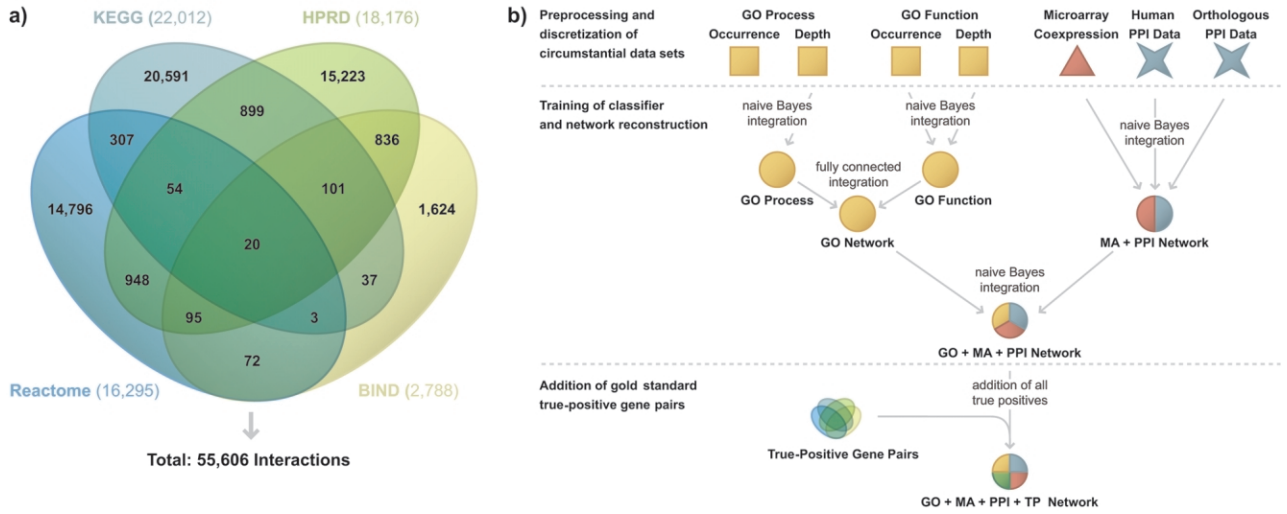

Figure 2 Integration of data sets in four gene networks. *a,* Data sets were benchmarked against a set of 55,606 known true-positive gene pairs derived from BIND, KEGG, HPRD, and Reactome and 800,608 true-negative gene pairs derived from GO. The Venn diagram indicates the data sources from which the true positives were derived and their degree of overlap. Numbers in parentheses indicate the number of interactions that are provided by each of the data sets. *b,* Potential gene-gene interactions derived from GO, microarray coexpression data, and human and orthologous protein-protein interaction data were integrated using a Bayesian classifier. The steps involved in building this classifier are shown.

mutual information was calculated between all the genes represented within each data set[33] if there were at least 10 nonmissing data points. As a preprocessing step, expression levels were ranked; this invertible reparameterization did not affect the mutual information. Next, for each pair of genes, the joint distribution of expression levels was estimated by calculating a histogram with overlapping windows. The range was divided into six windows, where each window extends to the center of the next window. The number of windows was chosen by optimizing the error rate for the mutual information derived from analytical probability densities.[33] In this way, each data point contributes to two windows, except at the extremities. Finally, on the basis of the resulting distribution, the mutual information (MI) between each pair of genes was calculated as $\mathrm{MI(A,B)} = H(\mathrm{A}) + H(\mathrm{B}) - H(\mathrm{A,B})$, where $H(X)$ is the information-theoretic Shannon entropy.[34] For each microarray data set, the MI score was binned. This allowed the subsequent Bayesian classifier to determine the likelihood ratio, indicating whether gene pairs within each bin contained an overrepresentation of truly interacting gene pairs. Once the likelihood ratios had been determined for each data set, a receiver operator characteristic (ROC) curve was constructed, and the area under the curve (AUC) was calculated. Data sets that had a minimal AUC of 0.59 were combined in a naive way—for each gene pair, the likelihood ratios were multiplied by each other, resulting in a final microarray coexpression likelihood ratio for each gene pair.

Two orthologous protein-protein interaction data sets from Lehner and Fraser[20] were used to supplement the GO and microarray coexpression data. One data set contained computationally predicted human protein interactions that had been physically mapped within Ensembl genes. The second data set contained a subset of these protein pairs, to which Lehner et al. had assigned a higher confidence. Three bins were constructed: one containing the higher-confidence gene pairs, one containing the remaining lower-confidence pairs, and a third containing all the other unobserved gene pairs.

A human Y2H protein-protein interaction data set from Stelzl et al.[19] was integrated by mapping the HUGO identifiers to Ensembl genes. Two bins were constructed: one containing the gene pairs for which a Y2H interaction was reported, and one containing all the other unobserved gene pairs.

Network Integration

The Bayesian classifier was employed to integrate the various binned types of data. We chose not to learn the Bayesian network structure from the data but to use a predefined Bayesian network structure, for which the conditional probabilities were determined by benchmarking the various data sets against the gold standard (fig. 2) (details provided in appendix A). We subsequently generated four gene networks. One network contained evidence for interaction based on the GO data (GO network). Another network contained evidence for interaction derived from integrating the microarray coexpression and predicted protein-protein interaction data in a naive way (MA+PPI network). A third network combined, in a naive way, the GO and MA+PPI networks (GO+MA+PPI network), and this was complemented with all known true-positive interactions in a final network (GO+MA+PPI+TP network). To relate interacting genes directly or indirectly, an all-pairs shortest path was calculated for each gene network.[35] This measure of the minimal path length between pairs of genes was used in the subsequent method to associate disease genes with each other.

Disease Analysis and Positional Candidate-Gene Prioritization

Prioritizer assesses whether genes residing within different susceptibility loci are close together within the gene network. This indicates that this method could also work with diseases for which only two loci have been identified. However, in such a case, there is a considerable probability that two genes, each residing in a different locus, would interact by chance. We therefore restricted the analysis to diseases for which at least three contributing disease genes had been identified. These diseases and disease genes were derived from the Online Mendelian Inheritance in Man (OMIM) database,[36] by text mining the first paragraphs of all OMIM disease entries as of March 1, 2005, and extracting the OMIM gene numbers contained within these paragraphs (table A1 in appendix A). The HUGO gene name was later extracted from these OMIM entries and was mapped to an Ensembl gene name. If, for any one disease, there were two disease genes situated at the same chromosome and positionally <200 genes apart, one of the two genes was randomly removed to ensure that no loci would overlap.

The diseases for which at least three disease genes remained after filtering were analyzed by artificially generating susceptibility loci around the disease genes, in a range from 50 to 150 genes, in steps of 50. All 20,334 genes were assigned an initial effect score of zero, and, subsequently, all loci were traversed. Using each gene network for all positional candidate genes residing in a particular locus, we determined whether any of these genes were functionally closely related to genes physically residing inside another susceptibility locus. If this was the case, the effect score of the related gene that was functionally close but physically in another locus increased (fig. 1), by use of the following Gaussian kernel scoring function:

$$\text{effect} = e^{\frac{-\text{distance}^2}{53}},$$

where "distance" is defined as the all-pairs shortest path between the two genes. The kernel function width was chosen arbitrarily, but a sensitivity analysis showed that different widths did not influence the results much (data not shown). By applying this function, positional candidate genes that resided in different loci but that were functionally closely related in the gene network were assigned higher scores than positional candidate genes that were functionally far apart from each other. To correct for differences in topology of the gene network, an empiric *P* value was determined for each positional candidate gene through permutation of the other loci 500 times by reshuffling them across the genome and recalculating the effect scores. This permitted a probability density function to be determined per positional candidate gene, for which the empiric *P* value could be looked up. For each locus, the positional candidate genes were prioritized on the basis of this *P* value.

Results

Construction of a Functional Gene Network

The basis for our human gene network was a gold standard of validated gene-gene interactions (true positives) and a further set of gene-gene pairs that were deemed highly unlikely to interact (true negatives). To construct the set of true-positive gene pairs, 2,788 confirmed, direct, physical protein-protein interactions were derived from BIND; 18,176 confirmed human protein interactions were derived from HPRD; 22,012 direct functional interactions were derived from KEGG; and 16,295 interactions were derived from Reactome. This resulted in 55,606 unique true-positive gene relationships (fig. 2*a*). For the true-negative set, gene pairs were selected that encode for proteins localized in different cellular compartments. The combinations of cellular compartments were selected from their underrepresentation in the set of true positives (see the "Material and Methods" section). This resulted in 801,108 pairs, of which 500 were known to be true-positive gene pairs, and these were therefore removed from the set of true-negative gene pairs.

We trained the classifier on this gold standard and constructed functional human gene networks on the basis of GO data, microarray coexpression data, and inferred protein-protein interactions, as well as combinations of these. First, for each gene pair, we assessed whether the genes shared GO annotations, which were derived for 15,045 genes from Ensembl. Sharing of GO terms was based on the frequency of the least-common GO term shared between two genes and the maximal depth in the GO hierarchy at which two shared terms lay. Gene coexpression was calculated in 186 microarray data sets derived from GEO and 75 data sets from SMD. However, most of these data sets were not highly informative, as judged by their ability to identify true-positive gene interactions with a low false-positive rate. Because it is known that many classifiers perform best when a subset of features are used,[37,38] we used only four informative microarray coexpression data sets for classification,[39–42] each showing a minimal AUC of 0.59. In total, these data sets contained 461 microarray hybridizations. Finally, protein-protein interactions were derived from the Lehner and Fraser[20] data set containing human protein interactions predicted by mapping physical protein interactions from various *Saccharomyces cerevisiae, Drosophila melanogaster,* and *Caenorhabditis elegans* interaction data sets to orthologous human gene pairs. Of the 71,806 predicted gene pairs, we were able to physically map 62,635 gene pairs with both genes in the pair mapping to known Ensembl genes. A subset was defined by Lehner and Fraser[20] that contained 10,652 gene pairs deemed to be of higher confidence, of which 10,139 gene pairs could be mapped. In addition, we used 3,186 human protein-protein interactions identified by automated Y2H interaction mating by Stelzl et al.,[19] of which 1,751 could be mapped to different Ensembl gene pairs.

We assessed the performance of our classifier on the

basis of these various data sources in three different gene networks generated on the basis of a Bayesian framework, after preprocessing and binning of the data sets. As mentioned above, one network was generated solely on the basis of GO data (GO network), one network was based on both microarray coexpression and predicted protein-protein interaction data (MA+PPI network), and an overall network contained all three types of data (GO+MA+PPI network). ROC curves (fig. 3) show the performance of the reconstructed GO, MA+PPI, and GO+MA+PPI gene networks, which were constructed by cross-validating all data sets 10 times against the gold standard set, to mitigate overfitting (details provided in appendix A). When we compared the performance of the various gene networks, it became evident that the GO data set provided the most accurate evidence for interaction. The AUC was 88%, compared with 50% for an uninformative classifier. The ROC for the MA+PPI network shows that coexpression data derived from microarray expression, in conjunction with the orthologous protein-protein interaction data, correctly inferred functional interactions (AUC = 68%), but to a lesser extent than the GO network. Nevertheless, as can be deduced from the GO+MA+PPI network, addition of the microarray coexpression and the orthologous protein-protein interaction data to the GO network improved slightly the accuracy of the network (AUC = 90%). In accordance with most networks described in the literature thus far,[43] our reconstructed networks have a connectivity that follows a scale-free power-law distribution, which has also been demonstrated for other organisms.[44–46] This is most apparent when the topology of the MA+PPI network is assessed (see appendix A).

To validate our network, we used a list of 2,574 Y2H interactions that recently became available[47] to assess whether our gene network had predicted an interaction for these gene pairs. We first mapped the set to Ensembl pairs and then removed all pairs that were in our gold standard true-positive set, to ensure that we only assessed newly identified interactions. This resulted in a set of 1,318 novel gene pairs.

We then assessed whether our gene network had predicted an interaction for these pairs. While Y2H interactions are known to regularly yield false-positive results,[48] we decided to test whether the distribution of likelihood ratios for these gene pairs was significantly different from a null distribution of 10,000 gene pairs sampled by generation of random gene pairs by selecting two genes at a time from the set of all individual genes that made up the Y2H gene pairs. The results show that the 1,318 Y2H gene pairs have a significantly higher likelihood ratio than the null distribution ($P = .0003$, by Wilcoxon Mann-Whitney test), which indicates that our gene network is capable of inferring as-yet-unknown interactions.

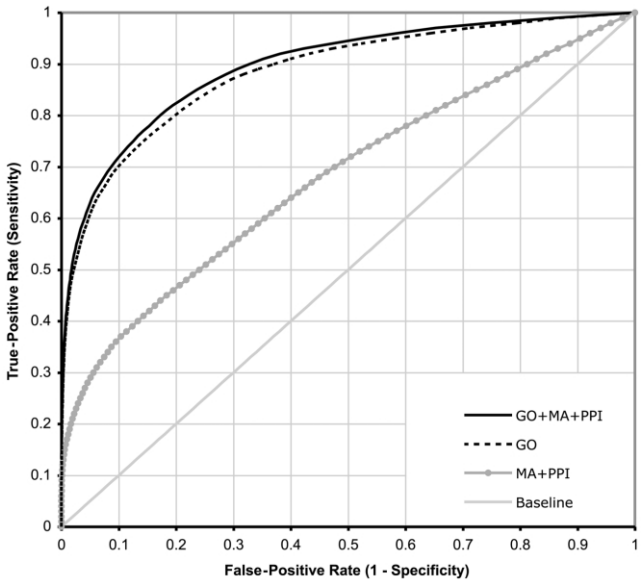

Figure 3 ROC curve of the GO network, the MA+PPI network, and the combined GO+MA+PPI network. The baseline (*solid gray line*) indicates the performance of a classifier that would be totally uninformative.

To allow researchers to look up known and predicted interactions and to identify the shortest routes between genes and susceptibility loci, we developed a Web tool, which is publicly available at the GeneNetwork Web site. The known and predicted interactions can be shown for each gene of interest, along with information about the source of evidence from which they were derived and how strong this evidence was. In addition, there are interactive graphs to visually explore how multiple genes interact with each other. All the data files (including the sets of true-positive and true-negative gene pairs) can be downloaded, along with a Java application programming interface, which can facilitate the development of new methods that use this gene network. Every 2 mo, we will update the gene network, on the basis of the most recent releases of the various repositories used in its construction.

Increased Functional Interactions Shown by Genes Associated with a Particular Disease

We first examined our hypothesis that genes associated with genetic disorders frequently share functional links, by assessing whether, for a disease, these causative genes were functionally more closely related to each other than a set of genes of equal size that were randomly selected from the full set of 345 unique disease genes of the 409 disease genes that were extracted from OMIM entries on disorders for which at least three causative genes were known. This set of disease genes was used

as a background distribution to prevent bias, since the disease genes are generally better characterized than the complete set of genes in the network. We generated one extra network (GO+MA+PPI+TP network) that complemented the GO+MA+PPI network with all known true-positive gene pairs, and we calculated the shortest direct or indirect distance between all pairs of genes. In 76 (79%) of the 96 diseases, the total distance between all combinations of disease genes in one disease was, on average, lower than the total distance between all combinations of randomly selected disease genes in 10,000 permutations. This confirms our hypothesis that, in the majority of diseases, the causative genes are indeed closely related functionally.

Genes implicated in disease processes tend to be studied more than those not implicated, which could result in a bias in the gene network based on GO annotations, since these represent known functional annotations. To assess the degree to which this possible bias affected our gene network, we looked at network connectivity. The average number of direct interactions involving disease genes was 199, compared with an average of 203 for the other 11,875 genes that interacted with at least one other gene. This indicates that other genes are equally represented in the gene network, despite the fact that disease genes may have been studied more.

Increased Power to Detect Disease Genes Provided by a Functional Gene Network

Usually, researchers pick a limited number of candidate genes in susceptibility loci to follow-up, because it is too costly and labor intensive to analyze all the genes residing in these loci. As a result, these studies have a limited chance of finding disease-related variants, largely depending on the size of the loci and the number of genes selected. Using a test set of known disorders in a similar setup, we evaluated the ability of our reconstructed network to correctly prioritize positional candidate genes in a set of top-ranked candidate genes of typical size (5–10 genes). The test set consisted of 96 different disorders, for which a total of 409 disease genes (345 unique genes) had been identified. These were obtained from OMIM, with 3–10 disease genes per disease (average 4.3 genes per disease) (table A1 in appendix A). Of the diseases, 59 are of Mendelian origin, 17 have complex inheritance, and 20 are various types of cancer (table 1).

The ability of the functional human gene network to correctly prioritize known disease genes was assessed by creating artificial, nonoverlapping susceptibility loci around these disease genes. Since many genes in these loci have no known or predicted interactions in our network, we only assessed those genes for which interactions were predicted, to prevent a bias toward genes that were better represented in the underlying high-throughput data sets. This resulted in susceptibility loci of varying widths, containing 50, 100, or 150 genes, which were predicted to interact with at least one other gene. If, for any particular disease, two disease genes residing in the same chromosome yielded loci that were partly overlapping, one of the two loci was randomly removed.

For each locus, the genes were traversed, and, for each gene, we assessed whether there was another gene residing in a different locus that was nearby within the gene network. The effect scores (see the "Material and Methods" section) of each gene were affected by the gene in the other locus that had the shortest path to that gene. This procedure has the potential to preferentially identify genes with many interacting partners over genes that are less well connected, because a highly connected gene has a higher chance of interacting with a gene residing in another locus than a gene for which only a few interactions have been predicted. To overcome bias in the method toward genes that are highly connected, we corrected for differences in the network topology by permuting the susceptibility loci for each disease 500 times across the genome.

After all positional candidate genes were ranked on the basis of this permuted score, the results (see fig. 4 and table 1) indicated that this method was able to identify many of the disease genes in the top 5 or top 10 genes per locus. As expected from the ROC curves of the various gene networks (fig. 3), the performance of the MA+PPI network proved to be the least powerful. Nevertheless, the number of correctly ranked genes was higher than would be expected to occur by chance (fig. 4*a* and 4*b*; indicated by baseline) for many of the susceptibility loci widths. When assessing susceptibility loci that contained, on average, 100 genes, we found 8% and 12% of the disease genes were contained within the top 5 and top 10 per locus, respectively, compared with the 5% and 10% we would expect to find by chance. A lack of predictive performance of the MA+PPI network explains why the ranking did not improve considerably when this network was used, as is evident from inspection of the ROC curves (fig. 4*c*), which show the proportion of disease genes and nondisease genes that are returned when different sizes of sets of top-ranked genes per locus are assessed. For 86 of the 345 unique disease genes within the MA+PPI network, no interactions were predicted. Hence, they were ranked low, the more so because the 49, 99, or 149 other genes, residing together with each disease gene in the constructed susceptibility loci, had been selected on the premise that they interacted with at least one other gene. The GO network performed considerably better; when we used it to assess susceptibility loci that contained, on average, 100 genes, we found 16% and 24% of the disease genes were contained within the top 5 and top 10

Table 1

Overview of the 96 Diseases Studied with Prioritizer and the Number of Disease Genes per Disorder That Ranked in the Top 10 Genes per Susceptibility Locus, With Locus Widths of 100 and 150 Genes

Disease Type and Disease	OMIM Number	No. of Genes	No. of Genes Ranking in Top 10 at Locus Width of 100 Genes				No. of Genes Ranking in Top 10 at Locus Width of 150 Genes			
			MA+PPI	GO	GO+MA+PPI	GO+MA+PPI+TP	MA+PPI	GO	GO+MA+PPI	GO+MA+PPI+TP
Mendelian inheritance (59 diseases):										
Achromatopsia 2	216900	3	0	0	1	0	0	0	0	0
Achromatopsia 3	262300	3	0	0	1	0	0	0	0	0
Adrenoleukodystrophy, autosomal neonatal form	202370	5	0	4	2	4	0	4	2	3
Amyloidosis VI	105150	3	1	1	0	1	0	0	0	1
Amyloidosis, familial visceral	105200	3	0	0	0	0	0	0	0	0
Amyotrophic lateral sclerosis 1	105400	5	0	1	0	0	0	1	0	0
Atypical mycobacteriosis, familial	209950	5	1	4	2	4	1	1	0	1
Autonomic control, congenital failure of	209880	5	0	0	1	1	0	0	1	1
Bardet-Biedl syndrome	209900	8	0	2	2	3	0	1	1	1
Bare lymphocyte syndrome, type II	209920	4	1	0	1	4	0	0	1	4
Cardiomyopathy, familial hypertrophic	192600	9	0	7	4	4	0	7	3	3
Cholestasis, intrahepatic, of pregnancy	147480	3	1	0	0	0	0	0	0	0
Cholestasis, progressive familial intrahepatic 1	211600	4	1	1	1	0	0	1	1	1
Complex I, mitochondrial respiratory chain, deficiency of	252010	5	0	5	4	5	0	3	3	3
Coumarin resistance	122700	4	0	0	0	1	0	0	0	0
Dementia, Lewy body	127750	3	1	0	0	0	0	0	0	0
Epidermolysis bullosa junctionalis, disentis type	226650	4	1	0	1	0	0	0	0	0
Epidermolysis bullosa of hands and feet	131800	4	0	1	2	1	0	0	1	1
Fanconi anemia	227650	6	1	0	1	6	1	0	0	6
Fundus albipunctatus	136880	4	0	1	0	3	0	1	1	2
Generalized epilepsy with febrile seizures plus	604233	3	2	2	2	1	1	0	1	0
Glutaricaciduria IIA	231680	3	3	2	3	3	3	1	3	3
Hermansky-Pudlak syndrome	203300	6	0	1	0	0	0	0	0	0
Hirschsprung disease	142623	6	1	0	0	1	1	0	0	1
Hydrops fetalis, idiopathic	236750	4	0	0	1	0	0	0	1	0
Hypercholesterolemia, familial	143890	6	0	2	3	1	0	1	2	1
Hypertrophic neuropathy of Dejerine-Sottas	145900	4	0	2	2	2	1	1	1	1
Hypokalemic periodic paralysis	170400	3	0	0	0	2	0	0	0	0
Ichthyosiform erythroderma, congenital, nonbullous, 1	242100	3	0	2	2	1	0	1	1	1
Immunodeficiency with hyper-IgM, type 2	605258	3	0	1	0	2	0	1	0	1
Immunodeficiency with hyper-IgM, type 3	606843	3	0	2	0	2	0	1	0	1
Kartagener syndrome	244400	3	1	2	2	1	0	2	2	1
Keratosis palmoplantaris striata I	148700	3	0	1	1	3	0	1	1	3
Laron syndrome, type II	245590	3	0	1	0	2	0	0	0	1
Leber congenital amaurosis, type I	204000	7	0	3	2	1	0	0	1	2
Leigh syndrome	256000	6	3	4	3	3	1	4	3	3
Leukoencephalopathy with vanishing white matter	603896	5	5	5	5	5	5	5	5	4
Maple syrup urine disease, type IA	248600	4	1	1	1	4	0	0	0	3
Maturity-onset diabetes of the young	606391	5	1	0	1	3	1	0	1	0
Myasthenic syndrome, congenital, fast-channel	608930	3	0	2	2	3	0	2	2	3
Myasthenic syndrome, slow-channel congenital	601462	3	0	2	2	1	0	2	2	2
Myoclonic dystonia	159900	3	0	0	0	1	0	0	0	1
Nemaline myopathy 1, autosomal dominant	161800	3	0	1	1	0	0	1	1	0
Nesidioblastosis of pancreas	256450	3	0	1	0	0	0	1	1	0
Night blindness, congenital stationary	163500	3	0	0	1	0	0	0	0	0

Obsessive-compulsive disorder 1	164230	3	0	1	0	0	0	0	0	0
Ossification of the posterior longitudinal ligament of spine	602475	3	0	0	0	1	0	0	0	0
Osteopetrosis, autosomal recessive	259700	3	1	0	0	0	0	0	0	0
Peters anomaly	604229	4	0	0	1	2	0	0	0	0
Pituitary dwarfism III	262600	3	0	0	0	2	0	0	0	1
Progressive external ophthalmoplegia	157640	3	1	1	0	0	0	0	0	0
Pseudohypoaldosteronism, type I, autosomal recessive	264350	3	0	2	2	1	0	2	2	0
Pulmonary alveolar proteinosis	265120	3	0	2	1	0	0	2	1	0
Refsum disease, infantile form	266510	3	1	1	1	2	0	1	1	2
Reticulosis, familial histiocytic	267700	3	0	0	0	0	0	0	0	0
Rhizomelic chondrodysplasia punctata, type 3	600121	3	0	1	0	2	0	1	0	1
Stickler syndrome, type I	108300	3	0	0	0	2	0	0	0	0
Waardenburg-Shah syndrome	277580	3	0	1	1	1	0	2	1	0
Zellweger syndrome	214100	8	1	4	4	7	1	3	4	5
Complex inheritance (17 diseases):										
Alzheimer disease	104300	8	0	1	0	3	0	1	1	2
Diabetes mellitus, non–insulin-dependent	125853	9	2	0	3	1	1	0	2	2
Elliptocytosis, Rhesus-unlinked type	130600	3	0	1	0	3	0	1	0	2
Graves disease	275000	3	0	1	0	1	0	0	0	0
Hypertension, essential	145500	7	1	1	0	0	0	0	0	0
Hypospadias	146450	3	0	0	0	1	0	0	0	1
IgA nephropathy	161950	4	1	0	0	1	1	0	0	1
Inflammatory bowel disease 1	266600	4	0	0	1	1	0	0	0	1
Longevity	152430	4	0	1	0	0	1	0	0	0
Lupus erythematosus, systemic	152700	4	0	0	0	0	0	0	0	0
Mycobacterium tuberculosis, susceptibility to infection by	607948	3	0	0	0	0	0	0	0	0
Myoclonic epilepsy, juvenile	606904	4	0	1	0	1	0	1	0	0
Obesity	601665	7	1	1	1	4	2	0	1	3
Osteoporosis, involutional	166710	5	0	1	1	0	0	3	1	2
Parkinson disease	168600	4	0	0	0	4	0	1	0	3
Rheumatoid arthritis	180300	5	0	0	0	0	0	0	0	1
Sudden infant death syndrome	272120	3	0	2	2	0	0	1	1	0
Heritable cancer (20 diseases):										
Bladder cancer	109800	3	0	0	0	0	0	0	1	0
Breast cancer	114480	10	2	1	4	2	1	0	2	1
Chondrosarcoma	215300	4	1	1	0	2	0	0	0	1
Esophageal cancer	133239	8	1	0	1	5	1	0	0	2
Glioma of brain, familial	137800	6	1	1	1	0	2	1	0	0
Hepatocellular carcinoma	114550	3	0	0	0	1	1	0	0	0
Juvenile myelomonocytic leukemia	607785	4	0	3	2	1	0	1	1	1
Leiomyoma, uterine	150699	4	0	0	1	0	0	0	0	0
Lung cancer	211980	4	1	0	1	2	0	0	1	0
Lymphoma, non-Hodgkin, familial	605027	4	0	2	2	2	0	1	1	2
Medulloblastoma	155255	4	1	0	2	2	1	0	1	0
Myeloma, multiple	254500	4	1	1	0	0	1	0	0	1
Osteogenic sarcoma	259500	3	0	0	0	0	0	0	0	1
Pancreatic carcinoma	260350	6	1	1	1	0	1	0	1	0
Pheochromocytoma	171300	3	0	0	0	0	0	0	0	0
Prostate cancer	176807	9	1	0	1	0	0	0	0	0
Renal cell carcinoma, papillary	605074	3	1	0	0	1	1	0	0	1
Rhabdomyosarcoma 2	268220	3	0	0	0	0	0	0	0	0
Thyroid carcinoma, papillary	188550	5	2	0	0	0	1	0	0	0
Turcot syndrome	276300	3	2	2	2	1	2	2	3	2
Total		409	49 (12%)	99 (24%)	93 (23%)	138 (34%)	33 (8%)	67 (16%)	68 (17%)	98 (24%)

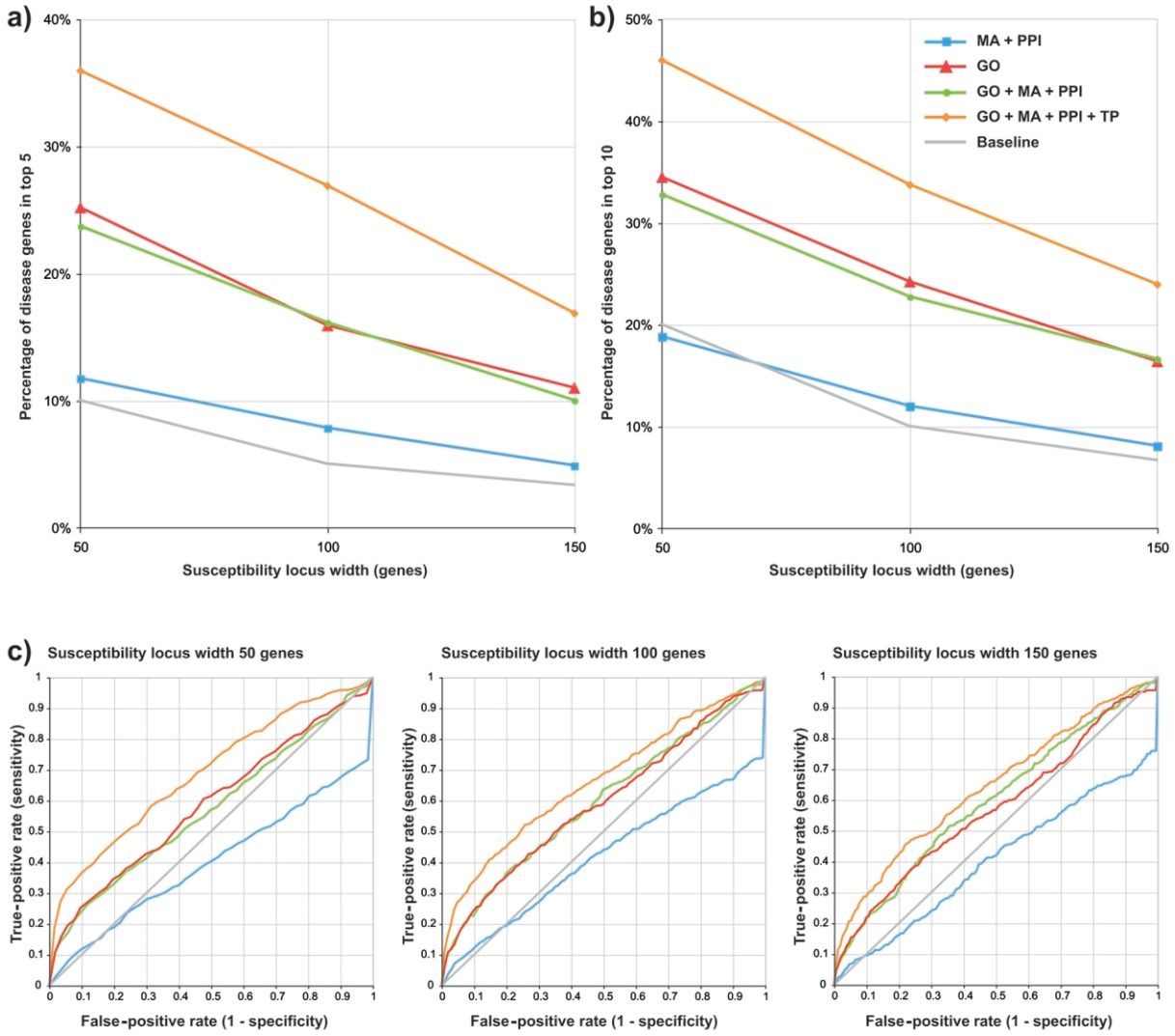

Figure 4 Accuracy of positional candidate-gene prioritization. *a* and *b*, Percentage of the 409 disease genes that was ranked among the top 5 (*a*) or top 10 (*b*) genes per locus, after artificial susceptibility loci of varying widths around these genes were constructed and when different types of gene networks were used. The baselines (*gray lines*) indicate the percentage of disease genes expected to rank among the top 5 or top 10 genes by chance. *c*, ROC curves for susceptibility loci that contain 50, 100, or 150 genes.

genes per locus, respectively. The performance of the disease analysis was best when the inferred GO+MA+PPI network was complemented with the known true-positive interactions (GO+MA+PPI+TP network); with this network and an average susceptibility locus width of 100 genes, 27% and 34% of the disease genes were contained within the top 5 and top 10 per locus, respectively.

We also assessed the probability of detecting at least one disease gene when only a fixed number of top-ranked genes per locus is followed up (fig. 5). When we employed the most comprehensive GO+MA+PPI+TP network and followed up all the top 5 or top 10 positional candidate genes for each disorder, using locus widths of 100 and 150 genes, we found at least one disease gene from these top sets of genes in 54% and 64% of the diseases, respectively, compared with 19% and 35% expected by chance. When we confined our analysis to diseases for which at least four or five disease genes were known, the performance of our method increased slightly (data not shown), because the true disease genes now interacted with more of the other true disease genes, increasing their overall scores.

Breast Cancer as an Example

We selected breast cancer as an example of how the various gene networks perform in a complex disease for which multiple disease genes have been identified. Ar-

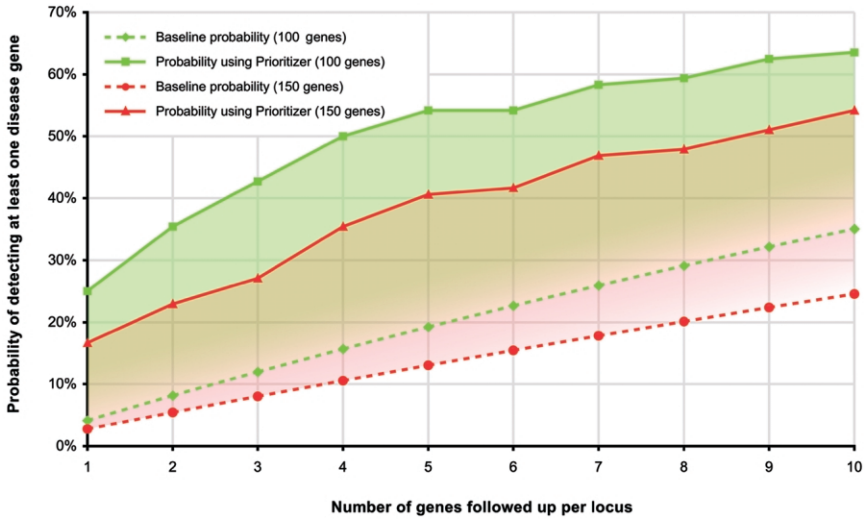

Figure 5 Probability of detecting at least one disease gene when a fixed number of top-ranked positional candidate genes—as ranked by Prioritizer—are followed up for each locus. Each locus contains either 100 or 150 genes, and the GO+MA+PPI+TP network was employed. The baselines (*dashed lines*) show the probability of detecting at least one disease gene if a fixed number of arbitrarily chosen genes in each locus are followed up.

tificial susceptibility loci, each comprising 100 genes, were constructed around 10 putative breast cancer genes described in OMIM (as of March 1, 2005). For each of the four networks, we then determined how many of the disease genes were ranked within the top 10 per locus. The MA+PPI network ranked two disease genes (*PIK3CA* and *CHEK2*) in the top 10, whereas the GO network ranked three (*BRCA2*, *NCOA3*, and *CHEK2*), and the GO+MA+PPI network ranked four (*BARD1*, *PIK3CA*, *TP53*, and *CHEK*) (fig. 6). However, the GO+MA+PPI+TP network, which integrates the most information, performed the worst; of the 10 disease genes now known, only 2 (*BARD1* and *BRCA1*) were ranked in the top 10. This can be explained by the observation that the true-positive set contained many known interactions for these 10 breast cancer genes. As the ranking procedure corrects for the topology of the network, these disease genes, with a marked increase in the number of relationships with other genes in this most comprehensive network, were suddenly no longer ranked as high. This became evident when the genes were ranked using the GO+MA+PPI+TP network but the differences in topology were not corrected for: 9 of the 10 breast cancer genes were then in the top 10 per locus.

Prioritizer Availability

To allow researchers to analyze susceptibility loci of interest, we developed a Java application that can be downloaded, along with regularly updated gene network definition files and source code from the Prioritizer Web site. After a set of susceptibility loci has been entered, Prioritizer ranks the positional candidate genes in each locus by using the method described above in conjunction with one of the four gene networks. It can generate two- and three-dimensional graphs of the top-ranked positional candidate genes, which allows the user to visually inspect how the genes within the different loci interact with each other.

Discussion

In this study, we describe the construction of a functional human gene network of considerable accuracy (fig. 3; AUC = 90%). As such, it can be used to assess interactions for a gene of interest through the bioinformatics tools that we have made available online. We have shown that, in cases where multiple genes underlie a disorder, these genes tend to have more functional interactions. When these functional interactions are employed to prioritize known disease genes in artificial susceptibility loci, the chance of detecting disease genes is increased considerably (2.8 fold).

In breast cancer, 4 of the 10 disease genes were ranked in the top 10 when the GO+MA+PPI network was applied, a fourfold enrichment over the single disease gene that would be picked up by chance. As has been discussed earlier, the correction for differences in topology is needed to prevent bias toward highly connected genes. However, this puts diseases in which underlying genes have a high degree of connectivity at a disadvantage, which was apparent in the analysis of breast cancer

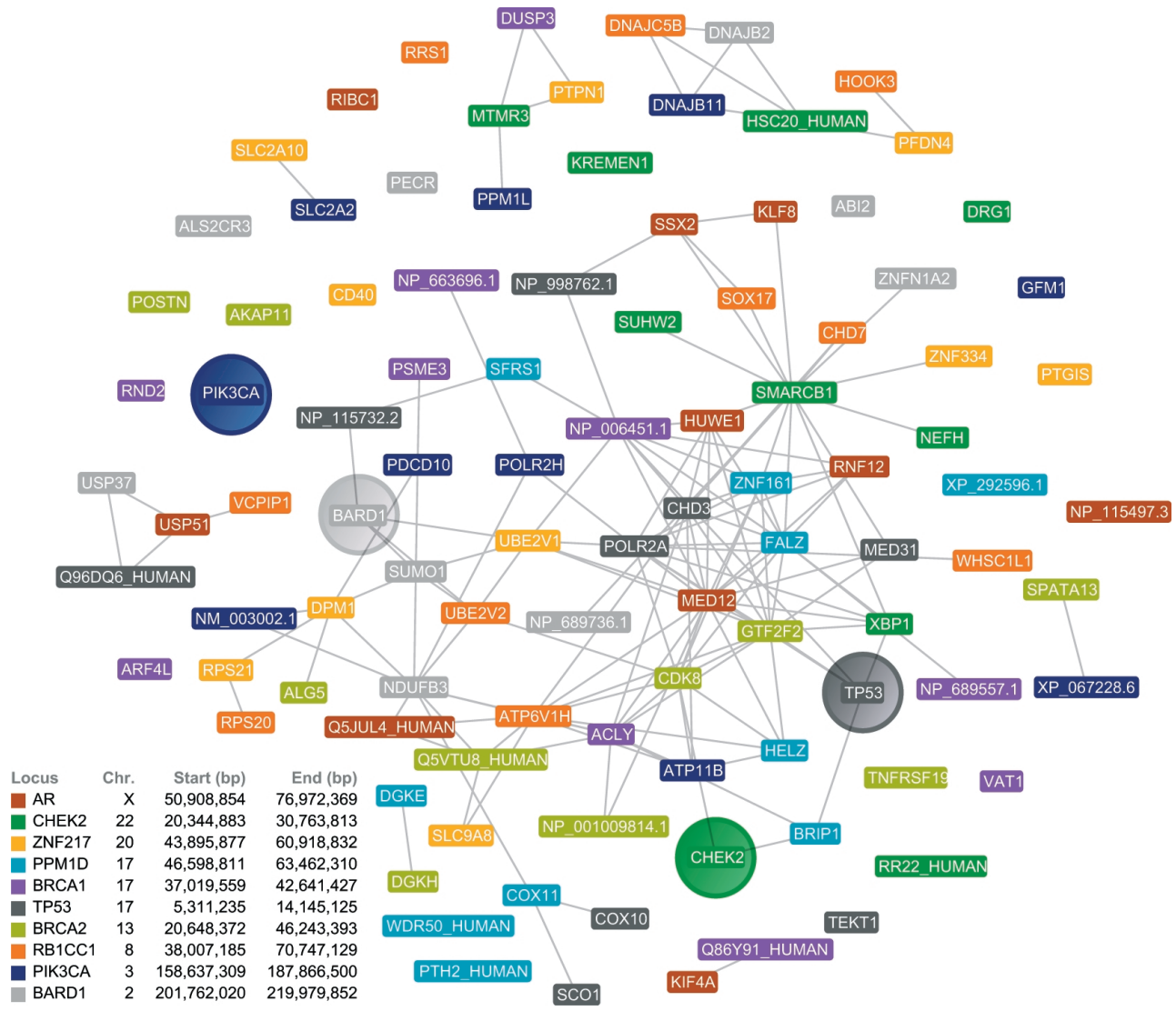

Figure 6 Prioritizer analysis of breast cancer. Susceptibility loci, each containing 100 genes, were defined around 10 known breast cancer genes. The 10 highest-ranked genes for each locus are shown in the graph, with colors indicating the locus in which they reside. Use of the GO+MA+PPI network led to four breast cancer genes (*PIK3CA*, *CHEK2*, *BARD1*, and *TP53* [*circles*]) being ranked in the top 10. Chr. = chromosome.

by use of the GO+MA+PPI+TP network. When this topology correction was omitted for breast cancer, the ranking of the disease genes improved considerably, to include 9 of the 10 genes. The availability of new high-throughput data sets will alleviate this problem in the future, by providing novel interactions for genes that currently have a low degree of connectivity, which will reduce the penalty on highly connected genes.

We noticed that the performance of Prioritizer was lower for complex disorders than for Mendelian disorders. This is likely caused by the fact that the etiology of complex diseases is more subtle and involves multiple pathways, so that most of the disease genes only confer a modest increased risk. Greater coverage of the gene network, leading to identification of relationships between genes that bridge the various pathways, could probably help to alleviate this problem.

When the accuracy of the various gene networks was assessed by investigation of their respective ROCs, it was envisaged that the GO+MA+PPI network would perform at least at a similar level in prioritizing disease genes as the GO network, because its AUC was greater. However, contrary to our expectation, when the positional candidate genes were prioritized, the disease genes in some diseases were ranked lower with the GO+MA+PPI network than with the GO network. One explanation

could be that, within the microarray coexpression data sets (the main contributor to the MA+PPI network), we did not distinguish between coexpression and coregulation. As such, many direct interactions between genes were inferred, but a large proportion of these interactions were actually indirect. Methods have recently appeared[33,49] that could help remove some of these incorrectly inferred interactions.

In a somewhat comparable method by Turner et al.,[21] positional candidate genes are prioritized by determining which genes share InterPro[50] domains and GO terms, as a measure to relate genes in susceptibility loci with each other. Our method extends this approach by also allowing for indirect relationships between individual disease genes, since Prioritizer uses the graph-theoretic distance between genes to relate them. Both approaches still rely largely on manual annotation, which is detrimental for genes that have not been investigated extensively. When no experimental evidence for interaction is available, there is only a small chance that these potential disease genes, residing in one specific susceptibility locus, will be associated with disease genes in other loci, since the sharing of GO or InterPro terms between these genes will be minimal. Although GO contributes the most to the performance of the Bayesian classifier, we should not depend entirely on a prediction if there is substantial evidence only from GO, while the evidence from the other data sets is lacking, for a specific gene pair, because the GO evidence has been inferred from the sharing of predominantly manually annotated terms, whereas the other sources rely more on direct biological measurements. It is expected that, when additional high-throughput data sets become available and their coverage of all possible functional interactions increases, GO evidence will be supplemented by experimental data, resulting in better predictions.

As such, an extensive and reliable functional gene network is crucial for good performance of our method. If this network is inaccurate or biased toward known genes, the ranking of true disease genes in the susceptibility loci will deteriorate. Several rapidly expanding data repositories are now becoming available that should help to improve our network. They include text mining methods,[51,52] which extract functional relationships from the literature, and methods that integrate results from high-throughput proteomic approaches.[53]

Our gene network, which, in its current form, has been applied to genetic linkage analysis, can also be used for other applications. Recently, efforts have been made to prioritize positional candidate genes on the basis of their expression,[54] with the assumption that differences in expression behavior in comparisons of patients with controls may be due to *cis*-acting variants in the underlying genes. However, it has turned out that, in most genes, differences in expression are determined by genetic variation in genes located elsewhere.[55,56] The reconstructed functional gene network can help to relate the observed differences in gene expression to the underlying causative genetic variants in other genes, which might help in identifying the disease genes.

Prioritizer might also be well suited for genomewide SNP association studies. Technical improvements in conjunction with decreasing costs now allow researchers to perform these studies in complex diseases, thereby considerably increasing the resolution at which one can assess genetic variation. However, as the number of tested SNPs increases, the number of tested individuals required to achieve sufficient power will also rise. To help overcome this problem, a new statistical method has recently been developed[57] that combines evidence from the most-significant tests, under the assumption that there are multiple true associations in the disease under investigation. However, within this confined set, the majority of genes will still be false positives because of power issues. Our positional candidate-gene prioritization method can easily be adapted to help distinguish true disease-associated genes and false-positive genes, by assuming that the true disease genes are mostly functionally related and will therefore be closer to each other in the gene network than to the false-positive genes that have been randomly selected.

We have demonstrated that it is feasible to use gene networks to prioritize positional candidate genes in various heritable disorders with multiple associated genes, even when the susceptibility loci are fairly large. As such, this article and the proposed methods show that the integration of gene networks with various genetic studies can be useful in identifying disease genes. We envisage that improvements both in the quality of the data sets making up these gene networks and in the statistical methods incorporating the networks will result in new, genetically testable hypotheses.

Acknowledgments

We thank Jackie Senior and members of the Complex Genetics Section and the Department of Human Genetics for critically reading the manuscript. This study was supported by Netherlands Organization for Scientific Research grant 901-04-219 and by a grant from the Celiac Disease Consortium, an innovative cluster approved by the Netherlands Genomics Initiative and partially funded by a Dutch government grant (BSIK03009).

Web Resources

The URLs for data presented herein are as follows:

Biomolecular Interaction Network Database (BIND), http://bind.ca/
Ensembl, http://www.ensembl.org/index.html
GeneNetwork, http://www.genenetwork.nl
Human Protein Reference Database (HPRD), http://www.hprd.org/

Kyoto Encyclopedia of Genes and Genomes (KEGG), http://www.genome.jp/kegg/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/
Prioritizer, http://www.prioritizer.nl
Reactome, http://www.reactome.org/

References

1. Jacobi FK, Broghammer M, Pesch K, Zrenner E, Berger W, Meindl A, Pusch CM (2000) Physical mapping and exclusion of *GPR34* as the causative gene for congenital stationary night blindness type 1. Hum Genet 107:89–91
2. Seri M, Martucciello G, Paleari L, Bolino A, Priolo M, Salemi G, Forabosco P, Caroli F, Cusano R, Tocco T, Lerone M, Cama A, Torre M, Guys JM, Romeo G, Jasonni V (1999) Exclusion of the *Sonic Hedgehog* gene as responsible for Currarino syndrome and anorectal malformations with sacral hypodevelopment. Hum Genet 104:108–110
3. Simard J, Feunteun J, Lenoir G, Tonin P, Normand T, Luu The V, Vivier A, et al (1993) Genetic mapping of the breast-ovarian cancer syndrome to a small interval on chromosome 17q12-21: exclusion of candidate genes *EDH17B2* and *RARA*. Hum Mol Genet 2:1193–1199
4. Tumer Z, Croucher PJ, Jensen LR, Hampe J, Hansen C, Kalscheuer V, Ropers HH, Tommerup N, Schreiber S (2002) Genomic structure, chromosome mapping and expression analysis of the human *AVIL* gene, and its exclusion as a candidate for locus for inflammatory bowel disease at 12q13-14 (IBD2). Gene 288:179–185
5. Walpole SM, Ronce N, Grayson C, Dessay B, Yates JR, Trump D, Toutain A (1999) Exclusion of *RAI2* as the causative gene for Nance-Horan syndrome. Hum Genet 104:410–411
6. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, et al (1994) A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. Science 266:66–71
7. Joenje H, Patel KJ (2001) The emerging genetic and molecular basis of Fanconi anaemia. Nat Rev Genet 2:446–457
8. D'Andrea AD, Grompe M (2003) The Fanconi anaemia/*BRCA* pathway. Nat Rev Cancer 3:23–34
9. de Winter JP, van der Weel L, de Groot J, Stone S, Waisfisz Q, Arwert F, Scheper RJ, Kruyt FA, Hoatlin ME, Joenje H (2000) The Fanconi anemia protein FANCF forms a nuclear complex with FANCA, FANCC and FANCG. Hum Mol Genet 9:2665–2674
10. Yamashita T, Kupfer GM, Naf D, Suliman A, Joenje H, Asano S, D'Andrea AD (1998) The Fanconi anemia pathway requires FAA phosphorylation and FAA/FAC nuclear accumulation. Proc Natl Acad Sci USA 95:13085–13090
11. Zatz M, de Paula F, Starling A, Vainzof M (2003) The 10 autosomal recessive limb-girdle muscular dystrophies. Neuromuscul Disord 13:532–544
12. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, et al (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res Database Issue 33:D418–D424
13. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, et al (2004) Human Protein Reference Database as a discovery resource for proteomics. Nucleic Acids Res Database Issue 32:D497–D501
14. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res Database Issue 32:D277–D280
15. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res Database Issue 33:D428–D432
16. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, et al (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res Database Issue 32:D258–D261
17. Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernandez-Boussard T, Jin H, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. Nucleic Acids Res Database Issue 33:D580–D582
18. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R (2005) NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Res Database Issue 33:D562–D566
19. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122:957–968
20. Lehner B, Fraser AG (2004) A first-draft human protein-interaction map. Genome Biol 5:R63
21. Turner FS, Clutterbuck DR, Semple CA (2003) POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol 4:R75
22. Brunner HG, van Driel MA (2004) From syndrome families to functional genomics. Nat Rev Genet 5:545–551
23. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, et al (2004) An overview of Ensembl. Genome Res 14:925–928
24. Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. Nat Rev Genet 5:251–261
25. Egmont-Petersen M, Feelders A, Baesens B (2005) Confidence intervals for probabilistic network classifiers. Comput Stat Data Anal 49:998–1019
26. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302:449–453
27. Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306:1555–1558
28. Xia Y, Yu H, Jansen R, Seringhaus M, Baxter S, Greenbaum D, Zhao H, Gerstein M (2004) Analyzing cellular biochemistry in terms of molecular networks. Annu Rev Biochem 73:1051–1087
29. Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29:131–163
30. Jansen R, Gerstein M (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr Opin Microbiol 7:535–545
31. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14:1085–1094
32. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11:1725–1729
33. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. Nat Genet 37:382–390
34. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–356
35. Floyd RW (1962) Algorithm 97: shortest path. Commun ACM 5:345
36. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a know-

ledgebase of human genes and genetic disorders. Nucleic Acids Res Database Issue 33:D514–D517
37. Jain A, Zongker D (1997) Feature selection: evaluation, application and small sample performance. IEEE Trans Pattern Anal 19:153–158
38. Waller WG, Jain AK (1978) Monotonicity of performance of Bayesian classifiers. IEEE Trans Inform Theory 24:392–394
39. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101:6062–6067
40. Stegmaier K, Ross KN, Colavito SA, O'Malley S, Stockwell BR, Golub TR (2004) Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. Nat Genet 36:257–263
41. Rieger KE, Hong WJ, Tusher VG, Tang J, Tibshirani R, Chu G (2004) Toxicity from radiation therapy associated with abnormal transcriptional responses to DNA damage. Proc Natl Acad Sci USA 101:6635–6640
42. Rieger KE, Chu G (2004) Portrait of transcriptional responses to ultraviolet and ionizing radiation in human cells. Nucleic Acids Res 32:4786–4803
43. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113
44. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, et al (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415:180–183
45. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, et al (2003) A protein interaction map of Drosophila melanogaster. Science 302:1727–1736
46. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147
47. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, et al (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437: 1173–1178
48. Vidalain PO, Boxem M, Ge H, Li S, Vidal M (2004) Increasing specificity in high-throughput yeast two-hybrid experiments. Methods 32:363–370
49. Egmont-Petersen M, de Jonge W, Siebes A (2004) Discovery of regulatory connections in microarray data. In: Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp 149–160
50. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, et al (2005) InterPro, progress and status in 2005. Nucleic Acids Res Database Issue 33:D201–D205
51. Yandell MD, Majoros WH (2002) Genomics and natural language processing. Nat Rev Genet 3:601–610
52. Malik R, Siebes A (2005) CONAN: an integrative system for biomedical literature mining. Lect Notes Artif Intell 3808:248–259
53. Bader JS, Chaudhuri A, Rothberg JM, Chant J (2004) Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 22:78–85
54. Franke L, van Bakel H, Diosdado B, van Belzen M, Wapenaar M, Wijmenga C (2004) TEAM: a tool for the integration of expression, and linkage and association maps. Eur J Hum Genet 12: 633–638
55. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:743–747
56. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297–302
57. Dudbridge F, Koeleman BP (2004) Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am J Hum Genet 75:424–435