Bioinformatics, 31(8), 2015, 1243–1249
doi: 10.1093/bioinformatics/btu799
Advance Access Publication Date: 4 December 2014
Original Paper

OXFORD

Systems biology

Graphical algorithm for integration of genetic and biological data: proof of principle using psoriasis as a model

Lam C. Tsoi[1,*], James T. Elder[2,3] and Goncalo R. Abecasis[1,*]

[1]Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA, [2]Department of Dermatology, University of Michigan, Ann Arbor, MI, USA, and [3]Ann Arbor Veterans Affairs Hospital, Ann Arbor, MI, USA

*To whom correspondence should be addressed.

Associate Editor: Igor Jurisica

Received on August 25, 2014; revised on November 9, 2014; accepted on November 26, 2014

Abstract

Motivation: Pathway analysis to reveal biological mechanisms for results from genetic association studies have great potential to better understand complex traits with major human disease impact. However, current approaches have not been optimized to maximize statistical power to identify enriched functions/pathways, especially when the genetic data derives from studies using platforms (e.g. Immunochip and Metabochip) customized to have pre-selected markers from previously identified top-rank loci. We present here a novel approach, called Minimum distance-based Enrichment Analysis for Genetic Association (MEAGA), with the potential to address both of these important concerns.

Results: MEAGA performs enrichment analysis using graphical algorithms to identify sub-graphs among genes and measure their closeness in interaction database. It also incorporates a statistic summarizing the numbers and total distances of the sub-graphs, depicting the overlap between observed genetic signals and defined function/pathway gene-sets. MEAGA uses sampling technique to approximate empirical and multiple testing-corrected *P*-values. We show in simulation studies that MEAGA is more powerful compared to count-based strategies in identifying disease-associated functions/pathways, and the increase in power is influenced by the shortest distances among associated genes in the interactome. We applied MEAGA to the results of a meta-analysis of psoriasis using Immunochip datasets, and showed that associated genes are significantly enriched in immune-related functions and closer with each other in the protein–protein interaction network.

Availability and implementation: http://genome.sph.umich.edu/wiki/MEAGA

Contact: tsoi.teen@gmail.com or goncalo@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In the past decade, association studies have been used to investigate genetic architecture of complex traits and diseases (McCarthy *et al.* 2008). Studies ranging from genotyping-based genome-wide association studies (GWAS; WTCCC 2007; Yang *et al.* 2010) to large-scale meta-analyses of GWAS (Liu *et al.* 2013; Teslovich *et al.* 2010), to the more recent next-generation sequencing-based association analysis (Tang *et al.* 2013; Zhan *et al.* 2013), have facilitated the identification of genetic susceptibility loci for different traits and diseases. However, identification of the underlying biological mechanisms from these powerful datasets has not always been straightforward.

To better understand these underlying biological mechanisms, it has been asked whether the identified loci contain genes enriched in

certain gene-sets such as genes annotated with biological functions [e.g. Gene Ontology: GO (Ashburner *et al.* 2000)] or pathways [e.g. KEGG (Kanehisa *et al.* 2012)]. It would also be possible to investigate if their encoding proteins could physically interact to form functional units such as heterodimeric cytokines (e.g. IL23A and IL12B as sub-units of cytokine IL23) or transcription factor complexes (e.g. p53). While advances in genotyping technology have greatly increased the power of available datasets, the development of optimal strategies to bridge between genetic and functional results remains a major theoretical and experimental goal of genetics research.

Depending on the genotyping platform being used, loci densities and genomic distribution of the markers in the association analysis might vary. One of the challenges in enrichment testing is to select the proper set of background markers/genes as 'null' observations to approximate its overlap against the gene-set in question. Previous studies have proposed different strategies to approximate the null observations: ALIGATOR (Holmans *et al.* 2009) randomly samples markers from the original platform to mimic its empirical marker distribution; INRICH (Lee *et al.* 2011) exhibits a sampling strategy using genomic interval as unit, and the sampled intervals are constrained to have similar marker and gene densities as those from the input intervals (i.e. susceptibility loci). Each of the sampled null observations from the above approaches would be used to construct empirical distributions for the numbers of overlap with functional gene-sets, and used to estimate the statistical significance of the observed data.

Both of the above count-based approaches enumerate the number of associated genes overlapping with each gene-set and work well (Holmans *et al.* 2009; Sklar *et al.* 2011) in identifying biological functions or pathways if the platform was designed in genome-wide scale (i.e. most common genetic markers are well-tagged by the genotyped markers). Recently, cost-effective genotyping platforms, such as the Immunochip (Cortes and Brown 2011; Parkes *et al.* 2013) and Metabochip (Voight *et al.* 2012), containing common and low-frequency variants have been designed to facilitate the identification of novel association signals (Beecham *et al.* 2013; Ellinghaus *et al.* 2013; Tsoi *et al.* 2012) and with the goal to fine-map established loci (Gong *et al.* 2013; Voight *et al.* 2012). These platforms contain dense markers in the top-ranked loci from previous studies (e.g. studies in autoimmune diseases for Immunochip; studies in metabolic and atherosclerotic/cardiovascular traits/diseases for Metabochip). Performing functional enrichment analysis on the signals identified using these customized platforms poses analytical challenges, as the pre-selected markers are enriched in potential or known susceptibility loci. The potential to identify the correct underlying biological mechanisms/pathways can be revealed only if sufficient gene-sets-overlapping genes from the significant loci are identified to provide enough statistical power. Moreover, as different studies have identified interactions between genes in disease susceptibility loci (Cotsapas *et al.* 2011; Rossin *et al.* 2011), a gene-set enrichment approach that incorporates independent information from biological interaction data could thus increase the power to identify functions or pathways crucial to the disease.

We present here a novel approach, called Minimum distance-based Enrichment Analysis for Genetic Association(MEAGA), to perform functional/pathway enrichment test for results from association studies. Instead of only enumerating the overlap between genes in the associated regions and genes annotated in a gene-set (i.e. count-based; Holmans *et al.* 2009; Lee *et al.* 2011), we use graphical algorithm and incorporates a statistic to measure the closeness between the overlapping genes in Steiner Tree(s) identified in a user-defined interaction database. Here we demonstrate in simulation studies that MEAGA is more powerful than a count-based approach when protein-coding genes in associated regions tend to be closer with each other in an interaction network. We also applied MEAGA to the results of a meta-analysis of psoriasis, a common inflammatory and hyperproliferative skin disease, using Immunochip datasets (Tsoi *et al.* 2012). Although the original study did not identify any enriched biological functions among the identified loci, MEAGA is able to reveal associated genes are closer with each other in the protein–protein interaction (PPI) network and are also significantly enriched in immune-related functions.

2 Methods

2.1 Overview

MEAGA tests the hypothesis that genes from the susceptibility loci in the trait/disease-associated function/pathway are closer with each other in the biological interactome than any other genes. The overview of MEAGA is illustrated in Fig. 1. MEAGA takes the markers used in the association analysis as input. Users will pre-specify the

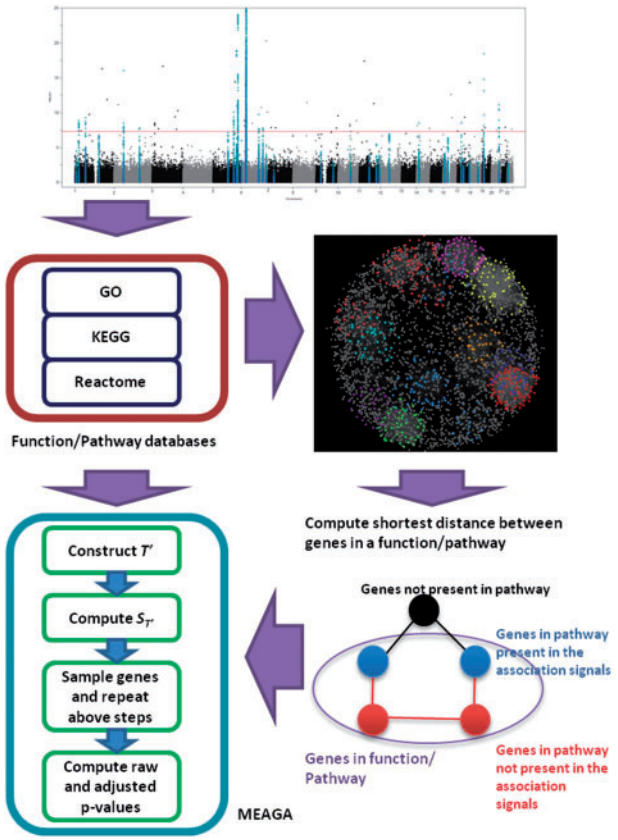

Fig. 1. The workflow overview of MEAGA. MEAGA performs enrichment analysis on association results using gene-sets obtained from functional/pathway databases. For each gene-set being tested, MEAGA identifies overlapping genes in association signals, and constructs Steiner Tree(s) in the biological interactome (e.g. PPI database) using Kou's algorithm. Kou's algorithm requires the shortest distances and paths between every gene-pair being computed among the overlapping genes. As illustrated in the bottom right figure, we could use genes present only in the gene-set (red lines in the bottom-right figure) or all genes (black lines) to compute shortest paths in the interactome. MEAGA then constructs Steiner trees and computes statistic S_T. MEAGA uses sampling technique to estimate empirical and multiple-testing corrected *P*-values

association signals (presumably the most significant marker in each of the significant loci) and annotate the tagged genes for each marker (e.g. using linkage disequilibrium blocks or genomic distance). For each functional gene-set being tested, MEAGA first identifies the overlapping genes from the signals, then uses graphical algorithms to construct sub-graph(s) with minimum distance(s) in the interactome. MEAGA next computes a statistic (S_T) summarizing the number of overlapping genes and the overall shortest distance(s) of the sub-graph(s). MEAGA uses a sampling strategy to approximate the null distribution of S and compute empirical and multiple testing-corrected P-values. There is no restriction for the type of interaction data [e.g. co-expression network (Langfelder and Horvath 2008); text-mining-derived networks (Franceschini *et al.* 2013); PPI network (Chatr-Aryamontri *et al.* 2013)] being used in MEAGA. Throughout this study, we used PPI data as this provides the experimentally verified information and it is well-known that proteins often work together for many biological functions (Cotsapas *et al.* 2011).

2.2 Data

We obtained the functional and pathway annotation data from the GO (Ashburner *et al.* 2000), KEGG (Kanehisa *et al.* 2012) and Reactome (Croft *et al.* 2013) databases (latest versions downloaded on May 7, 2013). Since the GO database has a directed acyclic graph (DAG) layout, in which each term is refined by one or more less-specific term in the same domain, we processed the GO's gene-to-GO file so we also annotated each gene with the 'ancestral' terms of its annotations in the DAG of GO database. Altogether, we obtained over 1.5 million gene to function/pathway connections from 18,987 genes and 19,383 functions/pathways. We downloaded and examined PPI data from three different sources: BioGrid (Chatr-Aryamontri *et al.* 2013), HPRD (Keshava Prasad *et al.* 2009) and STRING (Franceschini *et al.* 2013) (Supplementary Materials).

2.3 MEAGA

2.3.1 Sub-graphs with minimum distances

We first denote G' as the overlapping genes between a gene-set and genes in the susceptibility loci (G). By assuming genes involved in similar biological mechanisms associated with the traits or diseases would tend to be closer with each other in the interaction network (see Section 3), we use graphical distance as metric. Specifically, from the interactome network, we identify Steiner trees (T') among genes in G'; each Steiner tree is a connected and undirected acyclic graph with minimum total length. Note that a Steiner tree would allow other genes in the interactome to act as intermediate nodes when there is no direct connections between genes in G'. Therefore, if G' contains all the genes in the interactome Steiner tree problem would become the well-studied minimum spanning tree problem (Kruskal 1956; Prim 1957). We use the heuristic Kou's algorithm (Kou *et al.* 1981), which has been applied to other biological problems using Steiner tree technique (Richards *et al.* 2010; Zheng and Lu 2007), to identify T'.

2.3.2 Algorithm

MEAGA uses Kou's algorithm to identify Steiner tree(s) among genes in G' in the interactome for each of the function/pathway being tested. It is possible that not all genes in G' are connected in the interactome, so there could be more than one Steiner tree identified. Also, any gene in G' not presents in the interactome would be treated as an isolated Steiner tree. We developed a statistic which summarizes the number of genes in G' ($|G'|$), the number of identified Steiner trees ($|T'|$), and the distances for all T':

$$S_{T'} = \frac{|G'|\sum_{i\in T'} m_i w_i}{|T'|},$$

where m_i is the number of genes from G' in Steiner tree i, and w_i is the weight inversely proportional to the total distance (d_i) in tree i. Two Steiner trees could have the same m but different total distances if there are different numbers of intermediate nodes in the interactome coming from genes not in G'; similarly, two Steiner trees could have same total distance but contain different m values. Therefore, the shortest possible distance for a tree with size m could be used as a reference to normalize d. For Steiner tree with more than one gene, we set $w_i = o_i/d_i$, where o_i is the shortest possible distance for tree with size m_i, and for genes in G' isolated in the interactome, we set $w_i = 1/D_{\max}$, where $D_{\max}$ is a penalty score that equals the most distant length among all the pairwise shortest paths between genes in the interactome.

Note that $\sum_{i\in T'} m_i/|T'|$ in $S_{T'}$ is the average number of genes in G' among T', and w_i is a m_i-depending weight. If all genes in G' are isolated in the interactome, $S_{T'}$ would equal $|G'|/D_{\max}$, which is proportional to $|G'|$. Therefore, count-based functional enrichment approach is a special case of MEAGA if the interactome does not provide any information among genes in G' (see Section 3 from simulation studies).

Variables such as gene length, linkage disequilibrium structure, and minor allele frequency of the variant could affect the results of the functional enrichment test. To compute the empirical P-values for each gene set effectively, we use a sampling strategy adopted from previous study (Holmans *et al.* 2009). Briefly, MEAGA first randomly samples markers used in the association analysis until the number of sampled genes equal $|G|$ (user would be able to define the marker-to-gene region assignment, see below). It then applies the algorithm described above to identify gene-set overlapping genes among the sampled genes, and computes the S statistic using the identified Steiner tree(s). MEAGA repeats the sampling procedure N times to construct null distribution of S, and the empirical P-value is computed as the proportion of samples in N with statistic S as large as the $S_{T'}$ of the observed data. Multiple testing is then performed by sampling from the N samples a random gene set as 'observed data', and then computes its empirical P-values among all tested function/pathway gene-sets based on comparing its statistic with those obtained from bootstrapping the N samples. This procedure is performed M times, and the multiple testing-corrected p-value is computed as the proportion of the M samples with minimum empirical p-value (across the function/pathway gene sets) less than or equal to that from the observed data.

2.3.3 Implementation

We implemented MEAGA in Python, and used the graphical features available in the package NetworkX (Hagberg *et al.* 2008). Statistical computations were implemented in R (http://www.R-project.org). The construction of Steiner trees and the computation of the statistic S for each sampling step could be time consuming, and MEAGA supports multiprocessing for these procedures using the multiprocessing module in Python. The Kou's algorithm requires the shortest distances and their paths to be first computed between all G'. For effective performance, we pre-computed all shortest paths between genes in each function/pathway gene-set and stored them in a database to be readily retrieved when performing the Kou's algorithm. Figure 1 demonstrates that we could use genes present only in

a gene-set (red lines in the bottom-right figure) or all genes (black lines) to compute shortest paths in the interactome. For the identifications of Steiner trees and the calculations of statistic S in the analysis from this study, MEAGA would run linearly with the number of functions tested (e.g. 200 queried genes and 10 000 random gene-sets would take around 10 s to complete one function). The computation of adjusted P-values would take around 6 h to complete when N and M equal 10 000.

2.4 Simulation and application to association results

To evaluate the performance of MEAGA, and to compare the results with the count-based approach, we performed simulation to estimate the statistical power evaluated under the same type I error rate for each approach. We simulated null functions/pathways by selecting genes via randomly drawing markers successively from the set of all markers used in a real study (see below). The true associated functions/pathways were simulated similarly, except some of the genes were selected based on sampling the markers from the trait-susceptibility loci. We simulated shortest distances between the genes in the interactome, and varied the differences in distances for genes from the associated loci versus genes that are not. Statistical power is defined by the proportion of true associated functions identified to be significant. We used the 112 243 markers from a meta-analysis for psoriasis (2 Immunochip and 3 GWAS datasets; totally 10 588 psoriasis and 22 806 control samples) in the simulation to mimic real data (Tsoi *et al.* 2012).

Finally, we used MEAGA to perform functional/pathway enrichment test for the best single nucleotide polymorphisms identified in the 39 psoriasis independent loci. We used an interval of ±100 kb to define the tagging genes for each marker (Tsoi *et al.* 2012). Since functionally related genes could be from the same locus, multiple-counting this overlap would give rise to false positive results (Holmans *et al.* 2009; Lee *et al.* 2011). Therefore, we restricted our analysis to functions/pathways with genes in G' all coming from different loci.

3 Results

3.1 Simulation results

Figure 2 illustrates the results of the simulation studies under different approaches (count-based versus MEAGA), simulation settings (i.e. different numbers of total functions, trait associated functions, and genes in the functions), and weighting schemes used. We varied the distances of the Steiner trees identified from genes in association signals, using those obtained from randomly sampled genes as reference. As described above, the distance of a Steiner tree is inversely proportional to its weight w; the magnification scale '×' in Figure 2 is the relative weight set for associated genes referencing the non-associated genes in the platform. For example, '2×' in Figure 2 represents the distances for Steiner trees from randomly selected genes in the platform are set to be two times than those for associated genes.

Our results show that when the genes from association signals are relatively closer with each other in the interactome (i.e. 1.3× to 4×), MEAGA could obtain higher statistical power in identifying the true associated functions when comparing with count-based approach. The gain in statistical power increases when the relative weight of the Steiner trees increases for the associated genes. If the interactome does not provide any information (i.e. 1×), the count-based approach would become a special case for the MEAGA algorithm, and we illustrated in Figure 2 (black and grey lines) that the count-based approach has identical statistical power with MEAGA under this situation. We also simulated a 'counter intuitive' situation

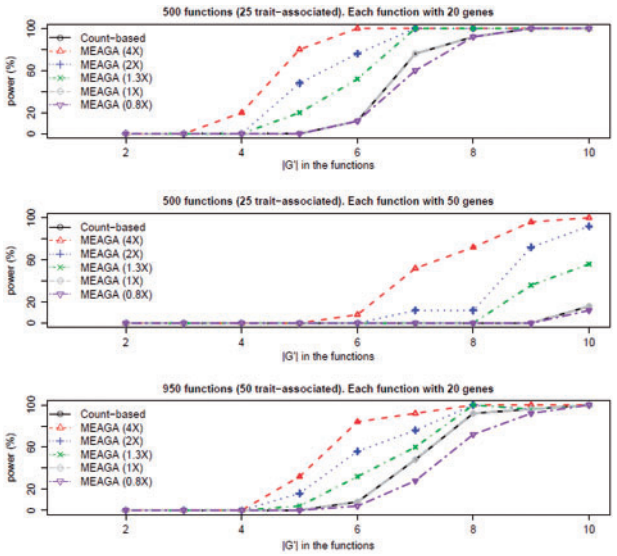

Fig. 2. Statistical power for identifying true functions associated with the trait in the simulation study by different approaches under different parameter settings. The x-axis represents the number of genes from the association loci overlapping with the genes in the function (i.e. $|G'|$). Different lines represent different relative weights used for associated genes, referencing those from non-associated genes. We simulated different numbers of total functions, trait-associated functions and genes in the functions for direct comparisons (top, middle and bottom panels). We also compared MEAGA to count-based approach (black). MEAGA uses sampling strategy adopted from ALIGATOR, so the count-based results equal to the results from ALIGATOR. Statistical power was evaluated at the significance threshold of $p_{adj} = 0.1$ criteria

where the associated genes are further away with each other in the interactome when comparing with the non-associated genes (0.8× in Fig. 2). The results show MEAGA only loses a small proportion of its power (purple in Fig. 2) when compared with the count-based approach. Section 3.2.1 illustrates that in reality the associated genes tend to be closer with each other in the interactome.

As expected, an increase in the number of gene-set overlapping genes (i.e. $|G'|$ in the x-axis of Fig. 2) would increase the statistical power in the identification of the true functions for all the scenarios being tested. The results also suggest the total number of genes in the functions (top and middle panels in Fig. 2) and the number of functions being tested (top and bottom panels in Fig. 2) both have effects on the statistical power. For example, when performing enrichment test for 500 small-sized (∼20 genes) functions/pathways (top panel in Fig. 2), we would gain the maximum power using different approaches if we have 10 overlapping genes from the association loci; however, testing for medium-sized (∼50 genes) functions/pathways (middle panel in Fig. 2) would only have less than 20% statistical power when using the count-based approach or if the interactome does not provide any information for associated genes, but we could achieve at least 80% statistical power if the total distances of Steiner trees from associated genes in the interactome are at least twofold shorter than those from non-associated genes. We also tested the performance of MEAGA when the gene interactions from the network are permuted to different degrees (Supplementary Fig. S1). While the results illustrate that the statistical power of accurately identifying associated functions/pathways decreases when the network is permuted to a greater degree, MEAGA would still perform as well as the count-based approach for a complete random network. In other words, when the interactome becomes

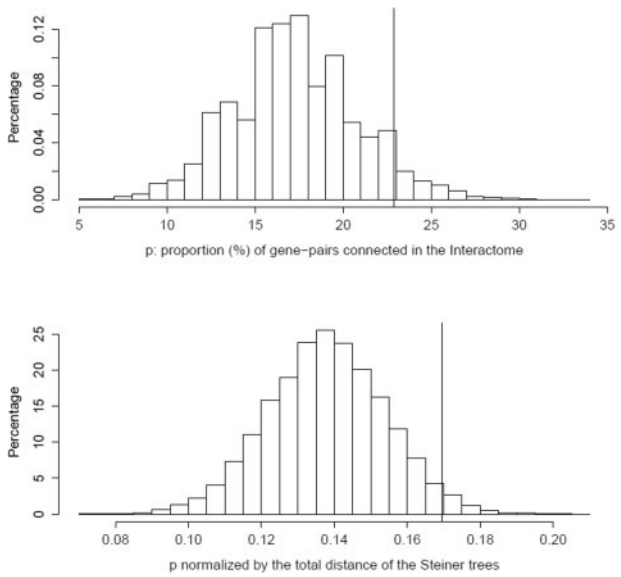

Fig. 3. (Top) Histogram for the proportion (p) of gene pairs connected (not necessarily directly connected) in the PPI network among 10 000 randomly sampled gene sets. The vertical line illustrated the proportion for genes from the associated loci. (Bottom) p normalized by the total Steiner trees' distances

uninformative, the count-based approach is a special case of MEAGA (i.e. it behaves like the 1× results in Fig. 2).

3.2 Application to meta-analysis results

3.2.1 Gene–gene connections in the interactome

Before performing functional enrichment testing on the meta-analysis results of psoriasis (Tsoi *et al.* 2012), we evaluated the connection for genes identified in the associated loci. We sampled 10 000 random gene-sets from markers used in the meta-analysis, and calculated the proportion (p) of gene pairs that are connected in the interactome as assessed by BioGrid (Fig. 3 upper panel). We found that this proportion for genes identified in the psoriasis susceptibility loci is among the top 5.5% (vertical line) of those from the randomly sampled gene-sets. We then normalized the proportion values by the total distances of the Steiner trees, and the normalized value of genes from the association loci is among the top 2.6% of the randomly sampled gene-sets (lower panel vertical line). In concordance with previous studies (Cotsapas *et al.* 2011; Rossin *et al.* 2011) suggesting that proteins encoded by genes in association loci identified in autoimmune diseases tend to be physically interact (via direct interaction or 2° of separation), our results show the first time that using the network topology (in terms of the distances of Steiner trees) of the interactome could enhance the resolution for identifying closely interacting genes from the associated loci in an autoimmune disease. The results further justify the assumption used in MEAGA of using information from the interactome (PPI in this case) for enrichment analysis to identify underlying biological functions/pathways. Performing the above analyses on PPI obtained from HPRD and STRING yielded very similar results (Supplementary Fig. S2).

3.2.2 MEAGA highlights immune-related functions/pathways for psoriasis

To perform enrichment testing on the best signals identified in the meta-analysis, we restricted our analysis to functions/pathways with at least 5 and at most 500 genes, and functions/pathways overlapping with at least two genes from the association signals. After limiting to the overlapping genes (G') coming only from different loci, we tested 743 functions/pathways using MEAGA. We first evaluated the impact of using different interaction data in the interactome on the results. The empirical P-values computed using only genes present in functions/pathways (x-axis) versus using all genes (y-axis) in the interactome are positively correlated (Spearman correlation = 0.88), and using only genes present in the functions/pathways tend to give stronger signals in the enrichment analysis (Supplementary Fig. S3). On further investigation, we realized that using all genes in the interactome to construct Steiner trees would tend to be less informative because the human PPI network is suggested to behave like a scale-free network (Barabasi and Oltvai 2004), with hub proteins lessening the average distances from one protein to the other. Therefore, using all interactions could introduce noise and thus MEAGA would behave like the count-based approach (spearman correlation = 0.94).

Using the genes present only in functions/pathways to construct the Steiner trees in the interactome specified by BioGrid, we identified 28 significantly enriched functions/pathways using the $p_{adj} \leq 0.1$ ($p_{raw} \leq 2.0 \times 10^{-4}$) criteria. Seven of these were identified as significant using ALIGATOR, whereas INRICH identified only one significant function (Supplementary Materials and Table S1). The top five results are shown in Table 1 (results from MEAGA using other PPI sources are illustrated in Supplementary Table S2). Sensitivity analysis using different parameters such as minimum/maximum number of genes allowed in the functions/pathways still yielded similar results. Immune-related functions such as 'regulation of leukocyte mediated cytotoxicity', 'positive regulation of lymphocyte differentiation' and 'regulation of myeloid cell differentiation' are significantly enriched among the genes from the psoriasis susceptibility loci. MEAGA also successfully reveals those biological functions (e.g. 'regulation of response to biotic stimulus') which contain three important genes forming protein complexes and involved in the IL-23 signaling pathways (i.e. *IL23A*, *IL23R*, *IL12B*) (Nair *et al.* 2009). Figure 4 demonstrates the Steiner trees identified in one of the enriched functions. While previous studies have suggested that TRAF3IP2 is one of the best candidate genes in the psoriasis susceptibility locus 6q21 (Ellinghaus *et al.* 2010), MEAGA shows that FYN from the same locus is annotated in an enriched immune-related function (Table 1 and Fig. 4), and in close proximity with genes from other susceptibility loci in the interactome. In fact, a recent study suggests FYN might be involved in the immunity in psoriasis, and could be a potential drug candidate (Manczinger and Kemeny 2013). Our results illustrate that MEAGA not only could identify enriched functions and interactions among associated loci, but also could prioritize the genes in the loci based on the functional and interaction data.

We also tested MEAGA for all the psoriasis susceptibility loci using markers from a GWAS study as background (Nair *et al.* 2009), and identified 95 significant functions highlighting different specialized immune-related pathways among the genes in the disease loci (Supplementary Table S3). We acknowledge that this analysis could be biased as the markers in the Immunochip datasets tend to be related to immune mechanisms, but it provides an overview of the biological functions or pathways that the genes from the psoriasis loci involved, when comparing with the enrichment in the whole genome in general.

4 Discussion

We present a novel functional enrichment method called MEAGA in this study. MEAGA uses graphical algorithm to identify significant

Table 1. Top 5 significant enriched functions identified by MEAGA for genes in psoriasis susceptibility loci

Significant functions	$\|G'\|$	$\|T'\|$	P	p_{adj}	G'
Regulation of response to biotic stimulus (87)	7	2	5.0×10^{-5}	2.6×10^{-2}	*DDX58, ELMO1, FYN, IL12B, IL23A, IL23R, TNFAIP3*
Regulation of transcription factor import into nucleus (67)	4	2	1.0×10^{-4}	4.0×10^{-2}	*DDX58, IL12B, IL23A, NFKBIA*
Regulation of leukocyte-mediated cytotoxicity (33)	6	3	1.5×10^{-4}	5.5×10^{-2}	*ICAM1, IL12B, IL23A, IL23R, NOS2, STAT5A*
Regulation of cell killing (36)	6	3	1.5×10^{-4}	5.5×10^{-2}	*ICAM1, IL12B, IL23A, IL23R, NOS2, STAT5A*
Positive regulation of lymphocyte differentiation (65)	5	2	2.0×10^{-4}	7.0×10^{-2}	*IL12B, IL23A, IL23R, IL4, STAT5A*

For illustration purpose, we restricted the significant results to those functions/pathways with at most 100 genes. Numbers in the bracket represent the number of genes annotated in the function. Full results shown in the Supplementary Materials. We used PPI obtained from BioGrid in this analysis.

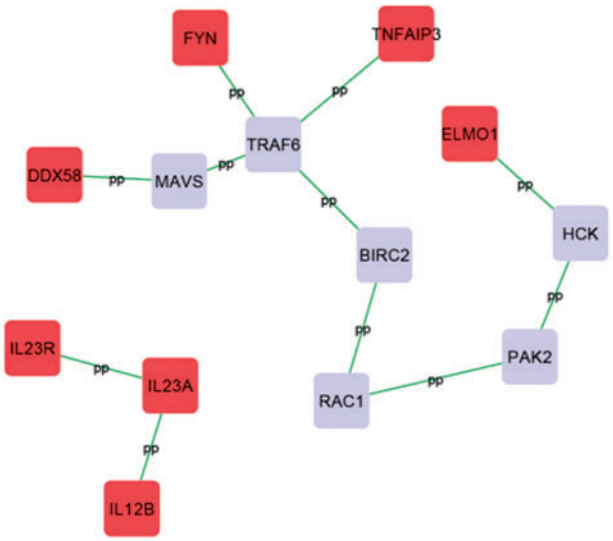

Fig. 4. The networks represent the Steiner trees identified in the enriched function 'regulation of response to biotic stimulus'. The edges are the PPI, and the nodes are the protein-coding genes in the function. Genes in dark grey colour are from the psoriasis susceptibility loci

biological functions containing genes from susceptibility loci which have high coherency in the interactome. We showed in the simulation studies that MEAGA outperforms count-based algorithm, and we also applied MEAGA to the results of a meta-analysis to identify immune-related functions potentially involved in the pathogenesis of psoriasis.

Systematic curation and interpretation of biological data are essential for making valid and novel biological inferences (Chatr-Aryamontri *et al.* 2013). In this study, MEAGA uses a systems biology approach to integrate functional annotations with PPI, and demonstrates that the genes from susceptibility loci of a complex disease are closer with each other in the PPI network than randomly chosen genes.

Extension to other sampling strategies is flexible for MEAGA, because its implementation distinguishes the Steiner Tree(s) construction from the permutation sampling step. We also demonstrated in the simulation studies the effectiveness of MEAGA given the same sampling strategy. Moreover, MEAGA supports predefined markers-to-gene-regions assignment, thus permitting flexibility for users to use either distance- (ALIGATOR) or linkage disequilibrium- (INRICH) based intervals when defining gene regions. Using MEAGA, we observed results similar to those shown above when using linkage disequilibrium-based intervals to define gene region for each associated marker (results not shown).

We have used graph theory techniques in this study, and we show that the systems biology approach could facilitate the understanding of biology for complex disease. Biological systems possess network properties and often function together, and there have already been several interesting published studies using network approaches to model biological units using genome-scale data such as gene expression or somatic mutation data (Alcaraz *et al.* 2014; Hofree *et al.* 2013; Ulitsky *et al.* 2010; Vandin *et al.* 2011). We believe advanced and efficient tools for providing biological inference of genetic data for complex traits could be based on these methodological foundations, and we have provided Supplementary Table S4 to illustrate how existing network algorithms which were applied to other genome-scale data could be useful for future studies of complex traits.

With the recent advancements in the technological and methodological aspects for using next-generation sequencing to perform genetic association tests, we envision that MEAGA may prove valuable in providing biological inferences for these studies. This is particularly so for cost-effective targeted sequencing-based studies (Tang *et al.* 2013; Zhan *et al.* 2013) when the investigated regions are pre-selected to have strong/suggestive association, integrating independent information would enhance the statistical power in identifying the underlying biological mechanisms of the trait/disease being studied.

Acknowledgements

The authors thank the insightful comments from Xiaowei Zhan, Christian Fuchsberger and our reviewers.

Funding

L.C.T. and J.T.E. are supported by National Institute of Health (NIH) grants R01 AR042742 and R01 AR050511, J.T.E. is also supported by NIH grants R01 AR054966, R01 AR062382, R01 AR065183 and by the Ann Arbor Veterans Affairs Hospital. G.R.A. is supported by research grants R01HG007022 from the National Human Genome Research Institute and R01EY022005 from the National Eye Institute.

Conflict of Interest: none declared.

References

Alcaraz,N. *et al.* (2014) KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with Cytoscape. *BMC Syst. Biol.*, 8, 99.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.,* **25**, 25–29.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev.,* **5**, 101–113.

Beecham,A.H. *et al.* (2013) Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.,* **45**, 1353–1360.

Chatr-Aryamontri,A. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.,* **41**, D816–D823.

Cortes,A.and Brown,M.A. (2011) Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.,* **13**, 101.

Cotsapas,C. *et al.* (2011) Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.,* **7**, e1002254.

Croft,D. *et al.* (2013) The Reactome pathway knowledgebase. *Nucleic Acids Res.,* **42**, D472–D477.

Ellinghaus,D. *et al.* (2013) High-density genotyping study identifies four new susceptibility loci for atopic dermatitis. *Nat. Genet.,* **45**, 808–812.

Ellinghaus,E. *et al.* (2010) Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2. *Nat. Genet.,* **42**, 991–995.

Franceschini,A. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.,* **41**, D808–D815.

Gong,J. *et al.* (2013) Fine mapping and identification of BMI loci in African Americans. *Am. J. Hum. Genet.,* **93**, 661–671.

Hagberg,A.A., Schult,D.A. and Swart,P.J. (2008) Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy 2008*. pp. 11–16.

Hofree,M. *et al.* (2013) Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115.

Holmans,P. *et al.* (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.,* **85**, 13–24.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.,* **40**, D109–D114.

Keshava Prasad,T.S. *et al.* (2009) Human protein reference database–2009 update. *Nucleic Acids Res.,* **37**, D767–D772.

Kou,L., Markowsky,G. and Berman,L. (1981) A fast algorithm for Steiner Tree. *Acta Informatica* **15**, 141–145.

Kruskal,J.B. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.* **7**, 48–50.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559.

Lee,P.H. *et al.* (2011) INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics (Oxford, England)* **28**, 1797–1799.

Liu,D.J. *et al.* (2013) Meta-analysis of gene-level tests for rare variant association. *Nature Genet.* **46**, 200–204

Manczinger,M. and Kemeny,L. (2013) Novel factors in the pathogenesis of psoriasis and potential drug candidates are found with systems biology approach. *PLoS One* **8**, e80751.

McCarthy,M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev.,* **9**, 356–369.

Nair,R.P. *et al.* (2009) Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat. Genet.,* **41**, 199–204.

Parkes,M. *et al.* (2013) Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev.,* **14**, 661–673.

Prim,R.C. (1957) Shortest connection networks and some generalizations. *Bell Syst. Tech. J.,* **36**, 1389–1401.

Richards,A.J. *et al.* (2010) Assessing the functional coherence of gene sets with metrics based on the Gene Ontology graph. *Bioinformatics (Oxford, England)* **26**, i79–i87.

Rossin,E.J. *et al.* (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genetics* **7**, e1001273.

Sklar,P. *et al.* (2011) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.,* **43**, 977–983.

Tang,H. *et al.* (2013) A large-scale screen for coding variants predisposing to psoriasis. *Nat. Genet.,* **46**, 45–50.

Teslovich,T.M. *et al.* (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713.

Tsoi,L.C. *et al.* (2012) Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat. Genet.,* **44**, 1341–1348.

Ulitsky,I. *et al.* (2010) DEGAS: de novo discovery of dysregulated pathways in human diseases. *PLoS One* **5**, e13367.

Vandin,F., Upfal,E. and Raphael,B.J. (2011) Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.,* **18**, 507–522.

Voight,B.F. *et al.* (2012) The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genetics* **8**, e1002793.

WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.

Yang,J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.,* **42**, 565–569.

Zhan,X. *et al.* (2013) Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nat. Genet.,* **45**, 1375–1379.

Zheng,B. and Lu,X. (2007) Novel metrics for evaluating the functional coherence of protein groups via protein semantic network. *Genome Biol.,* **8**, R153.