Construction and Analysis of Genome-wide SNP Networks

Yang Liu*, Jin Zhou†, Zhiping Liu‡, Luonan Chen‡ and Michael K. Ng§
*Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
Email: 08466246@hkbu.edu.hk
†School of Information Science and Engineering, University of Jinan, Jinan, Shandong, China
Email: ise_zhouj@ujn.edu.cn
‡Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China
Email: zpliu@sibs.ac.cn, lnchen@sibs.ac.cn
§Centre for Mathematical Imaging and Vision, Department of Mathematics, Hong Kong Baptist University, Hong Kong
Email: mng@hkbu.edu.hk

***Abstract*—The study of gene regulatory network (GRN) and protein protein interaction network (PPI) is believed to be fundamental to the understanding of molecular processes and functions in system biology and therefore, attracted more and more attentions in past few years. However, there is little focus about network construction in single nucleotide polymorphism (SNP) level, which may provide a direct insight into mutations among individuals, potentially leading to new pathogenesis discovery and diagnostics. In this paper, we present a novel method to mine, model and evaluate a SNP sub-network from SNP-SNP interactions. Specifically, based on logistic regression between two SNPs, we first construct a genome-wide SNP-SNP interaction network. Then by using gene information, selected SNP seeds are employed to detect SNP sub-networks with a maximal modularity. Finally to identify functional role of each SNP sub-network, its gene association network is constructed and their functional similarity values are calculated to show the biological relevance. Results show that our method is effective in SNP sub-network extraction and gene function prediction.**

I. Introduction

Rapid advancements in gene regulatory network (GRN) and protein-protein interaction (PPI) network make the putative functional connections hidden among genetic codes previously appear in front of us intuitively and orderly [1], [2]. In the post-genomic era, more and more attempts are being made to a system level understanding of biological organisms, which is viewed as an integrated and interacting network of genes, proteins and biochemical reactions which give rise to life [3]. Fruitful works [4], [5], [6], [7], [8], [9], [10], [11] provide us a better understanding of the structure and dynamics of cellular and organismal function from an integrating point of view. In this work, we construct a functional network at single nucleotide polymorphism (SNP) level, and conduct a system-wide association analysis of the SNP sub-networks, which may provide a new perspective into pathogenetic mechanisms of complex diseases.

SNP is a DNA sequence variation occurring when a single nucleotide - A, C, G, or T - differs at the same position between individuals [12]. SNPs are believed to result in differences between individuals, such as susceptibility to diseases [13]. They are considered as the most abundant and invaluable markers in human genome, that is a potential powerful tool for both of genetic researches and applications in practice [14], [15]. It has been observed that SNPs seldom act as simple functions to a single gene locus while performing multiple interactions and inheritance together.

Studying the functionality of SNP is of particular interest due to high data volume and the complexity of interactions. A variety of computational approaches have been developed and implemented for selecting a cohort of SNPs. One kind of them aims to identify a subset of SNPs that are assumed to independently have effect on a genotype decision. Horne and Camp [16] applied principal component analysis to evaluate multivariate SNP correlations to infer groups of SNPs in linkage disequilibrium (LD) and to establish an optimal set of group-tagging SNPs during an informative association analysis. Recently, a new integrative scoring system for prioritizing SNPs based on their possible deleterious effects within a probabilistic framework is proposed [17]. A new multi-objective optimization framework based on the notion of Pareto optimality for identifying SNPs that are both informative tagging and have functional significance is successfully applied for lung cancer study [18].

Another kind of those approaches is based on pairwise associations of SNPs. This approach is to select a set of SNP pairs such that each of them is highly interacted with the other. Schwender and Ickstadt [19] employed logic regression to identify SNP interactions explanatory for the disease status in case-control study, and proposed two measures for quantifying the importance of these interactions for classification. A global partitioning based on pairwise associations of SNPs is defined by Katanforoush et al. [20], and the pairwise allelic association of SNPs selected can describe various features of genomic variation, in particular for recombination in the hotspots regions. Wan et al. [21] proposed a novel learning approach (SNPRuler) based on the predictive rule inference to find disease-associated epistatic interactions. Liu et al. [22] applied a shrunken methodology to genome-wide SNP selection and based on the pairwise SNP-SNP interaction values to construct SNP networks.

2012 IEEE 6th International Conference on Systems Biology (ISB)
978-1-4673-4398-5/12/$31.00 ©2012 IEEE

Xi'an, China, August 18–20, 2012

In this paper, we first construct SNP networks based on their interactions, and then identify SNP-SNP interaction modules or sub-networks for functional analysis. A gene association network with each SNP sub-network is further constructed through the corresponding gene functionalities and regulatory relationships to identify the functional roles of SNP sub-networks. Specifically, genes whose SNPs located in the same sub-network are extracted and their pairwise similarity values are computed based on the literature vocabularies, i.e., Gene Ontology (GO) terms [23]. Gene pairs that have a similarity value larger than a defined threshold will be considered as highly functionally similar and will be connected in the gene association sub-network. Clearly in such a manner, it may bring us into a new perspective about gene network construction. Based on this study, we can make a relationship between SNPs and genes in systems biology perspective.

The rest of the paper is organized as follows: In "Methods" section, our network identification and computational modeling for SNPs networks are discussed. In "Results" section, we present experimental results to show how to construct SNPs networks and their gene association networks, and also demonstrate the effectiveness on analysis of complex diseases. In "Discussion" section, some concluding remarks are given.

II. METHODS

A. SNP-SNP Interactions

As we deal with data sets that have disease-trait samples, we consider to test epistasis using PLINK to detect SNP-SNP interactions (http://pngu.mgh.harvard.edu/purcell/plink) [24], whose focus is purely on whole genome association analysis of genotype/phenotype data. All pairwise combinations of SNPs can be tested, although this may not be desirable in statistical terms, it is computationally feasible for moderate datasets using PLINK. Input SNPs can be tested using a logistic regression model, which is based on allele dosage for each SNP, A and B, and fits the model in the form of

$$Y \sim b_0 + b_1.A + b_2.B + b_3.AB + e \quad (1)$$

The test for interaction is based on the coefficient of b_3, therefore we only considers allelic by allelic epistasis. We note in PLINK software that there is another fast algorithm called fast-epistasis which use collapsed 2-by-2 contingency table, and the computation can be greatly accelerated.

B. Network Structure Analysis

To better interpret and analyze available SNP-SNP interactions information, we construct and study a genome-wide SNP network. The network is represented as an un-directed, unweighted graph with each SNP as a node and each SNP-SNP interaction as an edge. If two SNPs are significantly interacted with each other under a predefined threshold, there will be an edge connecting between these two SNPs, otherwise not.

As a genome-wide SNP network is huge, it is better to find interesting sub-networks or functional modules for analysis. Detecting densely connected regions within themselves but sparsely connected with the rest of the network therefore play a vital role in revealing important principles of cellular organization and function. Here our goal is to find a group of SNPs in a sub-network sharing the common cellular interactions and responsible for certain genetic functions or pathways.

Suppose the genome-wide SNP network is denoted by $G = \{V, E\}$, where $V = \{v_1, v_2, ..., v_j\}$ is set of vertices and $E = \{e_{ij}\}$ is the set of edges. In our study, V is a set of unique SNPs and an edge e_{ij} is defined as a pair of vertices (v_i, v_j) denoting the direct connection between vertex v_i and v_j, i.e., there is an interaction between the ith SNP and the jth SNP. We denote A to be the adjacent matrix of G. If V_1 and V_2 are two disjoint node subsets of V, we further define

$$L(V_1, V_2) = \sum_{i \in V_1} \sum_{j \in V_2} A_{ij}.$$

For a given sub-network G, it contains both the node set and the edge set.

The goal of this work here is to address about the sub-network for a given SNP. We are motivated by two factors. Firstly, due to the complexity and modularity of SNP networks, it is more feasible computationally to study a sub-network containing a small number of SNPs of interest. Secondly, sometimes the whole structure of the network may not be our primary concern. Rather, we may be more interested in finding the sub-network which contains a set of related SNPs (a set of related corresponding genes) of interest.

Our aim is to discover sub-networks such that SNPs inside the sub-network interact significantly and, meanwhile, they are not strongly influenced by SNPs outside the sub-network. Sub-networks are constructed starting with seeds consisting of one or more SNPs believed to be participated in a viable sub-network. For instance, seeds are the suspected SNPs that are related to a particular disease. The sub-network iteratively keeps its compaction by evaluating connectivity degree of each node in the remaining whole network, this node will be adjoined the sub-network if it is influenced more by inside than by outside and will be abandoned otherwise. A literature search is carried out in "Gene" database of NCBI with a search phrase of this particular disease name. Genes that have genotype data will be further filtered out. Then, genes will be manually annotated and we select some of them that have a clear relationship descriptions with diseases. An individual list of SNPs corresponding to each of these selected genes are downloaded and checked about overlaps with our own data set. For each gene, we pick out one or more SNPs that have a comparable larger connections with others as the seeds. Even through this list is probably not a complete one, it provides a good reference related to this particular disease.

In our algorithm, the sub-network detection is based on the modularity optimization. A sub-network is initiated by seeds and it keeps growing based on the maximization of a modularity calculation [25] shown as follows:

$$d(G_i) = \frac{L(V_i, V_i) - L(V_i, \overline{V}_i)}{|V_i|} \quad (2)$$

The sub-network detection procedure iteratively adds one more

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

vertex with the highest d value from all neighbors of existing sub-network. The initial sub-network in our method is a dense region constituting with all seeds. It is considered as the origin or the core, that can expand itself to get the greatest density of a community finally. We remark that a breadth-first spreading procedure is done during our method. Breadth-first search can find shortest paths from a single vertex v_i to all others in time $O(m)$. The spreading area is defined as all neighbors of the existing sub-network. For every vertex v_i in spreading area, we compute its d value, and this vertex will be admitted as one member of this sub-network if its d value is the highest among all neighborhoods and otherwise not. A detailed description is shown in Table I.

TABLE I
ALGORITHM DESCRIPTION.

Input: $G(V, E)$ is the whole graph with vertex set V and edge set E, $G^{'}(V^{'}, E^{'})$ is the sub-graph with vertex set $V^{'}$ and edge set $E^{'}$. $S = \{s_1, s_2, ..., s_n\}$ is the seed set.

1 $G^{'} := S$
2 for $\forall v_i \in G$, $v_i \notin G^{'}$, if $\exists v_j \in G^{'}$ that $A_{ij} = 1$
3 $d_{v_i} = \frac{L(V^{'} \cup v_i, V^{'} \cup v_i) - L(V^{'} \cup v_i, V - V^{'} \cup v_i)}{\left|V^{'}\right| + 1}$
4 where $L(V^{'}, V^{''}) = \sum_{i \in V^{'}} \sum_{j \in V^{''}} A_{ij}$
5 and A is adjacent matrix of G
6 end for
7 $V^{'} = V^{'} \cup v_i$ where $d_{v_i} = d_{max}$
8 Until $d_{v_i} < d_{max}$

III. EXPERIMENTAL RESULTS

A. Datasets

The Parkinson disease (PD) SNPs data is based on a genome-wide genotyping of 270 individuals with idiopathic Parkinson Disease cases (case) and 271 neurologically normal controls (control) downloaded from the Coriell Institute for Medical Research. The genotyping was performed using the Illumina Infinium I and Infinium II assays. The Illumina Infinium I assay asseses 109,365 unique gene-centric SNPs while the Infinium II assay assesses 317,511 haplotype taggings SNPs based upon Phase I of the International HapMap Project. The Illumina Infinium I and II assays share 18,073 SNPs in common. Therefore, the combination of the two assays after preprocessing represents 408,787 unique SNPs. A frequency and genotyping pruning was done before experiment by using PLINK [24]. After frequency and genotyping pruning, there are 377,833 SNPs.

As we focused on coding SNPs which cause a functional impact on the genes in this study, SNPs that are located in the gene area were filtered out using SNP Function Portal [26]. There were 184,452 remaining SNPs and these coding SNPs would be the input of our proposed method in the following study.

B. Genome-wide SNP Network

For these 184,452 SNPs, there are 17,011,177,926 unique pairs. Due to the intensive computation requirement, we adopted parallel computing to construct this network. All the 184,452 SNPs were divided into 27 subsets, while the first 26 subsets contained 7,000 SNPs each and the last subset only 2,452 SNPs. Each of the 27 subsets would be calculated SNP-SNP interaction value with all the other remaining subsets, so there will be $\sum_{i=1}^{27} i = 378$ combinations. We performed this huge computational work on a parallel computing cluster constituting with 378 CPU nodes, where each node is under the configuration of 4D, 3.0GHz processor, 1GB RAM, 120G hard disk drive storage and Windows XP operating system. The average running time for each CPU node was around 32 hours and the average file size generated by each node was around 3.5G.

TABLE II
STATISTICS OF INTERACTIONS & UNIQUE SNPS AT DIFFERENT THRESHOLDS IN PARKINSON DISEASE STUDY.

Threshold	Total number of Interactions	Total number of SNPs	SNP with Max Connectivity	No. of Edges
1×10^{-5}	45672	52685	rs7909279	30
5×10^{-5}	319624	152520	rs7909279	40
			rs3739776	40
			rs6560142	40
1×10^{-4}	718788	169865	rs4752071	54

In order to work on the most significant SNP-SNP interactions, different thresholds on P-values (1×10^{-5}, 5×10^{-5} and 1×10^{-4}) were considered. For example, when the threshold was set to be 1×10^{-5}, only the SNP-SNP interaction with being smaller than 1×10^{-5} would be studied, otherwise, the corresponding interaction would not be considered. Table II shows the number of interactions and the number of SNPs in the genome-wide SNP network under different thresholds. The column in "SNP with Max Connectivity" indicates the SNP ID whose interactions with others is the maximum among all SNPs in the network, and "No. of Edges" tells us how many SNPs they are connected in the SNP network.

C. Seed Information

According to the seed selection criterions described in "Network Structure Analysis" section, for Parkinson disease, there are 220 related genes found after a literature searching, where 120 of them have genotype data. We collected 17 genes that definitely have clear descriptions with Parkinson disease. An individual list of SNPs corresponding to each of these 17 genes were downloaded and checked about overlaps with our own data set. For each gene, we picked out one or more SNPs that have a comparable larger number of edge connections with other SNPs as the seeds for different thresholds. Table III shows a detailed information about these 32 seeds, which will be used as the initial seeds.

D. SNP Sub-networks Construction

In order to preserve the characteristics of SNP sub-networks under three different thresholds, resulting sub-networks obtained under the setting with a small threshold will be employed as initial seeds of sub-networks with a large threshold, e.g., a sub-network determined under the threshold of 1×10^{-5}

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

TABLE III
SEEDS INFORMATION IN PARKINSON DISEASE STUDY.

Seed	Gene ID	Gene Symbol	NO. of Edges 1×10^{-4}	5×10^{-5}	1×10^{-5}
rs3738814	23400	ATP13A2	14	6	0
rs4680	1312	COMT	19	9	0
rs1544325	1312	COMT	17	8	0
rs3758653	1815	DRD4	9	3	3
rs11246226	1815	DRD4	12	4	0
rs10205801	26058	GIGYF2	18	12	0
rs10211596	26058	GIGYF2	18	9	5
rs2199503	2932	GSK3B	16	9	0
rs10878247	120892	LRRK2	20	13	4
rs11564173	120892	LRRK2	11	7	0
rs11564203	120892	LRRK2	27	12	0
rs11829088	120892	LRRK2	25	14	0
rs874250	4729	NDUFV2	11	4	1
rs11660603	4729	NDUFV2	11	5	1
rs705316	4885	NPTX2	20	10	0
rs834835	4929	NR4A2	6	2	0
rs483366	5071	PARK2	29	21	2
rs2022988	5071	PARK2	17	14	6
rs4288183	5071	PARK2	32	8	0
rs9458583	5071	PARK2	29	11	0
rs161802	11315	PARK7	7	2	0
rs178932	11315	PARK7	11	6	0
rs650616	65018	PINK1	16	7	0
rs1043424	65018	PINK1	11	6	0
rs1884082	12	SERPINA3	12	1	0
rs8007632	12	SERPINA3	22	18	5
rs464049	6531	SLC6A3	16	7	0
rs11133767	6531	SLC6A3	19	9	2
rs356168	6622	SNCA	13	5	0
rs2736990	6622	SNCA	13	6	0
rs4242202	6620	SNCB	8	3	0
rs10517003	7345	UCHL1	11	5	0

was employed as an initial sub-network under the threshold of 5×10^{-5}, and then the algorithm was applied to this initial sub-network to detect a suitable SNP sub-network. And also, the sub-network gained when 5×10^{-5} would be continually used as the initial seeds when 1×10^{-4}. We remark that the initial seed is always contained in this hierarchical sub-network structure.

On the other hand, we would like to evaluate the quality of the resulting SNP sub-networks for three different thresholds. We employed the modularity definition proposed by Li et al. [25], which is called D value. The property of modularity of D suggests a basic topological concept during network analysis. A module in a network is a region with dense internal connectivity and sparse external connectivity. Li et al. defined this modularity in a quantitative manner for evaluating the partition of a network into communities based on the concept of average modularity degree. This D value can improve the resolution limit by considering the information on the number of nodes in a detected module and the total number of links in the whole network.

$$D=\sum_{i=1}^{m} d(G_i)=\sum_{i=1}^{m}\frac{L(V_i,V_i)-L(V_i,\overline{V}_i)}{|V_i|} \qquad (3)$$

This measurement evaluates the quality of SNP sub-networks. The larger of the D value, the better of the sub-network.

The modularity D value of our method under the thresholds of 1×10^{-5}, 5×10^{-5} and 1×10^{-4} are 14.12, 41.85 and -8.91 respectively. Results show that our method can get a comparable higher modularity under the threshold of 5×10^{-5},

TABLE IV
CHARACTERISTICS OF SNP SUB-NETWORKS IN PARKINSON DISEASE STUDY.

Seeds	1×10^{-5} NO. of SNPs	NO. of Edges	d-value	5×10^{-5} NO. of SNPs	NO. of Edges	d-value	1×10^{-4} NO. of SNPs	NO. of Edges	d-value
rs3738814	1	0	0.00	14	22	2.07	25	43	-0.52
rs4680	1	0	0.00	12	14	0.15	57	79	-0.69
rs1544325	1	0	0.00	19	21	0.25	48	62	-1.71
rs3758653	4	3	1.50	10	9	1.10	34	34	-0.74
rs11246226	1	0	0.00	14	15	1.14	26	34	0.00
rs10205801	1	0	0.00	18	46	3.50	22	84	3.27
rs10211596	7	6	1.57	15	16	1.00	36	45	-1.11
rs2199503	1	0	0.00	16	25	1.75	34	50	-1.21
rs10878247	5	4	1.40	22	21	0.86	46	52	-0.93
rs11564173	1	0	0.00	17	22	0.71	46	68	-0.96
rs11564203	1	0	0.00	14	13	0.21	40	45	-1.75
rs11829088	1	0	0.00	17	20	0.76	36	50	-0.97
rs874250	9	8	1.67	12	15	1.08	24	68	1.54
rs11660603	9	8	1.67	12	15	1.08	24	68	1.54
rs705316	1	0	0.00	12	16	1.33	24	41	-0.83
rs834835	1	0	0.00	12	11	0.42	34	34	-1.35
rs483366	9	10	2.11	31	64	2.19	52	120	0.17
rs2022988	9	8	1.67	26	37	1.73	46	125	1.02
rs4288183	1	0	0.00	7	6	0.43	21	23	-1.71
rs9458583	1	0	0.00	17	33	2.94	41	120	2.02
rs161802	1	0	0.00	6	5	0.67	34	59	-0.82
rs178932	1	0	0.00	6	5	0.33	31	43	-0.10
rs650616	1	0	0.00	16	17	1.00	38	78	0.55
rs1043424	1	0	0.00	12	26	2.92	22	52	0.64
rs1884082	1	0	0.00	7	8	1.43	20	24	-1.45
rs8007632	5	4	1.20	18	45	2.61	26	72	0.77
rs464049	1	0	0.00	10	9	0.20	27	27	-1.85
rs11133767	3	2	1.33	19	20	1.16	48	75	-0.79
rs356168	1	0	0.00	8	15	2.75	16	30	0.19
rs2736990	1	0	0.00	8	15	2.75	16	30	0.19
rs4242202	1	0	0.00	3	2	0.33	11	10	-1.45
rs10517003	1	0	0.00	4	3	1.00	8	7	0.13

which indicates that our sub-networks have a more organized structure in this circumstance. In fact, for most of the seeds, their SNP sub-networks have the highest d value when the threshold is 5×10^{-5}. Table IV shows the total number of SNPs and their edge connections in these 32 SNP sub-networks. We remark that as we are not interested in a complete partition of the whole network but the sub-networks with the seed SNPs.

E. SNP Sub-networks Annotation

For each SNP position, we can find out the associated gene in the chromosome. Based on the SNP sub-network, we can further construct the gene network of their associated genes based on their functions. Table V shows the detailed information of these gene association networks in terms of their number of genes and the number of edges. In addition, the similarity between two genes can be computed based on their molecular functions and biological process in Gene Ontology (GO) [23], which was implemented with a R Bioconductor package GOSemSim (http://www.bioconductor.org).

After a biological interpretation of these genes, we found that some of the gene association networks are directly or indirectly related to Parkinson disease. We chose one of the networks, growing from seed rs11133767 (gene locus is SLC6A3), as an example to demonstrate the effectiveness of our method. Subgraphs of a, b and c in Figure. 1 give the SNP sub-networks for three different thresholds and subgraphs of d, e and f in Figures. 1 give their corresponding associated gene networks. We can see from the gene association networks of both 1×10^{-5} and 5×10^{-5} that, two reported Parkinson related genes, PARK2 and SLC6A3 can be mined out by our algorithm and they can be connected directly, which provides a strong proof of the validity and feasibility of our method and also can be used to identify the functions of the respective SNP sub-networks. The other genes involved in this network

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

TABLE V
CHARACTERISTICS OF GENE ASSOCIATION NETWORKS AT 5×10^{-5} IN PARKINSON DISEASE STUDY.

Seeds	Gene Symbol	$\sum$ Similarity	NO. of Genes	NO. of Edges
rs3738814	ATP13A2	0.470	4	6
rs4680	COMT	3.138	10	45
rs1544325	COMT	4.528	12	66
rs3758653	DRD4	2.066	8	28
rs11246226	DRD4	2.425	8	28
rs10205801	GIGYF2	0.635	4	6
rs10211596	GIGYF2	1.709	6	15
rs2199503	GSK3B	1.259	7	21
rs10878247	LRRK2	9.860	18	153
rs11564173	LRRK2	2.659	10	45
rs11564203	LRRK2	3.050	11	55
rs11829088	LRRK2	3.497	12	66
rs874250	NDUFV2	0.252	4	6
rs11660603	NDUFV2	0.252	4	6
rs705316	NPTX2	1.045	6	15
rs834835	NR4A2	1.856	9	36
rs483366	PARK2	7.411	14	91
rs2022988	PARK2	3.358	12	66
rs4288183	PARK2	0.468	5	10
rs9458583	PARK2	1.184	6	15
rs161802	PARK7	0.560	4	6
rs178932	PARK7	0.659	5	10
rs650616	PINK1	2.941	10	45
rs1043424	PINK1	0.736	6	15
rs1884082	SERPINA3	0.239	9	36
rs8007632	SERPINA3	1.488	9	36
rs464049	SLC6A3	1.477	8	28
rs11133767	SLC6A3	3.506	11	55
rs356168	SNCA	0.119	2	1
rs2736990	SNCA	0.119	2	1
rs4242202	SNCB	0.015	2	1
rs10517003	UCHL1	0.495	4	6

TABLE VI
FUNCTIONS OF GENE ASSOCIATION NETWORK DERIVED FROM RS11133767 IN PARKINSON DISEASE STUDY.

Gene Symbol	Gene ID	Function	GO Terms
PARK2	5071	Mutations cause Parkinson disease and autosomal recessive juvenile Parkinson disease	35
SLC6A3	6531	Variation is associated with idiopathic epilepsy, susceptibility to Parkinson disease	28
FLJ33718	285489	Essential for neuromuscular synaptogenesis, functions in aneural activation of muscle-specific receptor kinase. Be a cause of familial limb-girdle myasthenia autosomal recessive	4
BTBD14A	138151	NACC family member 2, BEN and BTB (POZ) domain containing	7
PALLD	23022	Be associated with a susceptibility to pancreatic cancer type 1, also a risk for myocardial infarction	8
FLJ10292	55110	Mago-nashi homolog B	6
LRPAP1	4043	Low density lipoprotein receptor-related protein associated protein 1	20
GLP2R	9340	Stimulates intestinal growth and upregulates villus height in the small intestine	7
ADAMTS15	170689	ADAM metallopeptidase with thrombospondin type 1 motif	7
GRID2	2895	Be the predominant excitatory neurotransmitter receptors in the mammalian brain Play a role in neuronal apoptotic death	22
ERBB4	2066	Be activated by neuregulins and other factors and induces a variety of cellular responses including mitogenesis and differentiation. Mutations in this gene have been associated with cancer	33

are also thought to be indirectly related to Parkinson disease, for example, FLJ33718 is associated with myasthenia, GRID2 plays a role in neuronal apoptotic death and mutations in ERBB4 have been associated with cancer. These results share some features for Parkinson disease. The detailed function descriptions of these gene products and their number of GO terms discovered within Gene Ontology [23] at 5×10^{-5} are shown in Table VI.

IV. CONCLUSION

In this paper, we have presented a novel method to mine, model and evaluate SNP sub-networks from SNP-SNP interactions, which is further annotated by its respective gene association network. The SNP interaction of the proposed approach is based on logistic regression between two SNPs, by which we can construct a genome-wide SNP-SNP interaction network. We tested the proposed method for one real data set: Parkinson disease data. Some useful SNP seeds relevant to diseases were employed to detect SNP sub-networks with a maximal modularity. Their associated gene association networks were considered afterward and their functional similarity values were calculated to show the biological relevance. We found that some of the gene association networks reveal strong structural and functional relationships with diseases.

All in all, our framework can discover sub-networks within a whole-scale genome-wide network efficiently and can provide a new insight into the relationships between SNPs and genes. On the one hand, from SNP to gene level, the gene relationships expressed by SNP networks can be considered as an extension of NCBI, Gene Ontology or other biomedical databases. On the other hand, from gene to SNP level, existing mature gene network can help us modify or annotate SNP sub-networks, which can give a better explanation of their behavior in biological function and explore some potential functional relationships from SNP level to gene level.

ACKNOWLEDGMENT

We thank the participants and the submitters for depositing samples at the NINDS Neurogenetics repository. The samples for this study are derived from the NINDS Neurogenetics repository at Coriell Cell Repositories. Access to the samples and to these data are available from the website: http://ccr.coriell.org/Sections/BrowseCatalog when we first attempted to download it. Currently, it is available from NCBI: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gap. This work was financially supported by Research Grant Council [201812] and Hong Kong Baptist University FRGs.

REFERENCES

[1] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.-H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. jun Pan, M. J. Schilstra, P. J. C. Clarke, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood, and H. Bolouri, "A Genomic Regulatory Network for Development," *Science*, vol. 295, no. 5560, pp. 1669-1678, Mar. 2002.

[2] H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman, "Evolutionary Rate in the Protein Interaction Network," *Science*, vol. 296, no. 5568, pp. 750-752, Apr. 2002.

[3] H. Kitano, "Systems Biology: A Brief Overview," *Science*, vol. 295, no. 5560, pp. 1662-1664, Mar. 2002.

[4] M. Gustafsson, M. Hornquist, and A. Lombardi, "Constructing and Analyzing a Large-Scale Gene-to-Gene Regulatory Network-Lasso-Constrained Inference and Biological Validation," *ACM/IEEE Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 3, pp. 254-261, Jul. 2005.

[5] Y. Huang, I. M. Tienda-Luna, and Y. Wang, "A Survey of Statistical Models for Reverse Engineering Gene Regulatory Networks," *IEEE Signal Process Mag.*, vol. 26, no. 1, pp. 76-97, Jan. 2009.

[6] D. R. Rhodes, S. A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A. M. Chinnaiyan, "Probabilistic model of the human protein-protein interaction network," *Nat Biotechnol.*, vol. 23, no. 8, pp. 951-959, Aug. 2005.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

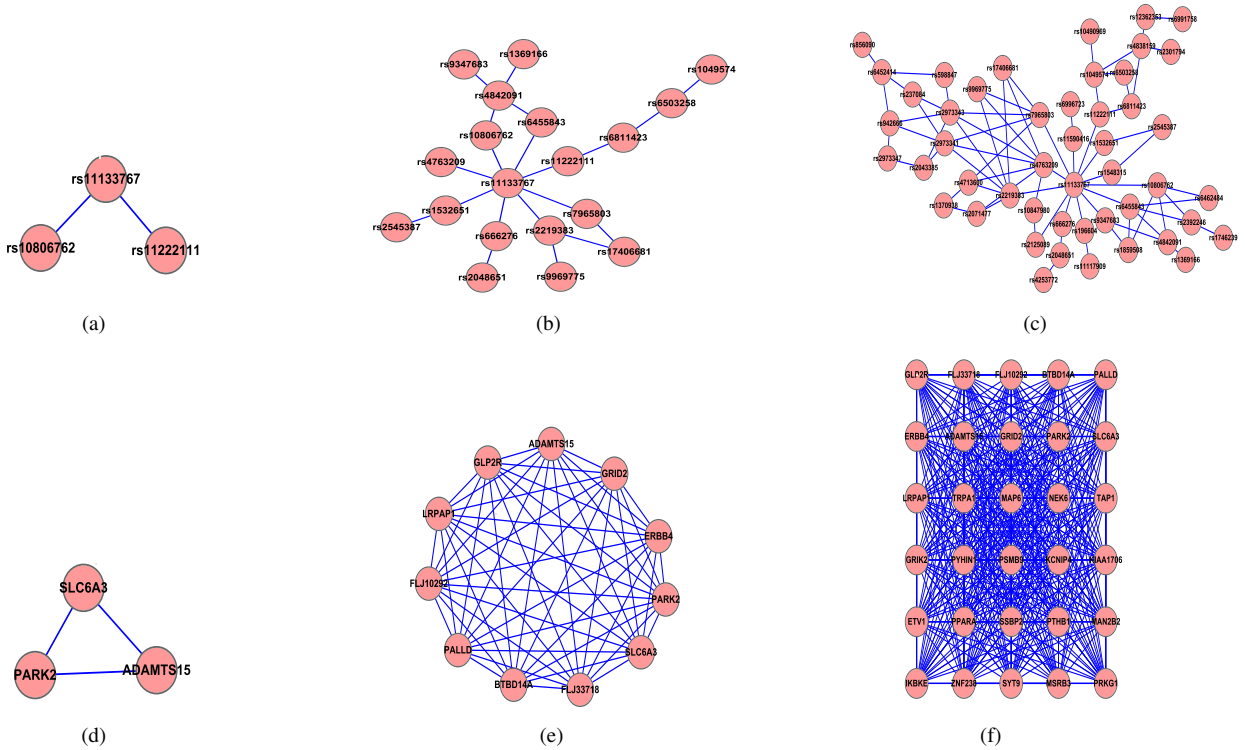

Fig. 1. SNP sub-networks of rs11133767 at different thresholds and corresponding gene association networks of PD data. The top three subgraphs are SNP sub-networks of rs11133767 at 1×10^{-5} (a), 5×10^{-5} (b) and 1×10^{-4} (c), where the bottom three subgraphs are the gene association networks derived from corresponding SNP sub-networks at 1×10^{-5} (d), 5×10^{-5} (e) and 1×10^{-4} (f).

[7] C. B. Huang, F. Morcos, S. P. Kanaan, S. Wuchty, D. Z. Chen, and J. A. Izaguirre, "Predicting Protein-Protein Interactions from Protein Domains Using a Set Cover Approach," *ACM/IEEE Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 1, pp. 78-87, Jan. 2007.

[8] A. Zhang, *Protein Interaction Networks Computational Analysis*, Cambridge University Press, pp. 1-294, 2009.

[9] A. Ikin, C. Riveros, P. Moscato, and A. Mendes, "The Gene Interaction Miner: a new tool for data mining contextual information for protein-protein interaction analysis," *Bioinformatics*, vol. 26, no. 2, pp. 283-284, Dec. 2009.

[10] L. Chen, R. S. Wang, X. S. Zhang, "Biomolecular Networks: Methods and Applications in Systems Biology", *Wiley, 2009, Malden MA*.

[11] L. Chen, R. Q. Wang, C. Li, K. Aihara, "Modelling Biomolecular Networks in Cells: Structures and Dynamics", *Springer-Verlag, 2010, London*.

[12] A. J. Brookes, "The essence of SNPs," *Gene*, vol. 234, no. 2, pp. 177-186, Jul. 1999.

[13] N. Risch, and K. Merikangas, "The future of genetic studies of complex human diseases," *Science*, vol. 273, no. 13, pp. 1516-1517, Sep. 1996.

[14] N. J. Schork, D. Fallin, and J. S. Lanchbury, "Single nucleotide polymorphisms and the future of genetic epidemiology," *Clin. Genet.*, vol. 58, pp. 250-264, Jul. 2000.

[15] J. N. Hirschhorn, and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nat. Rev. Genet.*, vol. 6, pp. 95-108, Feb. 2005.

[16] B. D. Horne, N. J. Camp, "Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation", *Genet Epidemiol.*, vol. 26, pp. 11-21, 2004.

[17] P. H. Lee, H. Shatkay, "An integrative scoring system for ranking SNPs by their potential deleterious effects", *Bioinformatics*, vol. 25, pp. 1048-1055, 2009.

[18] P. H. Lee, J. Y. Jung, H. Shatkay, "Functionally informative tag SNP selection using a pareto-optimal approach: playing the game of life", *BMC Bioinformatics*, vol. 10, 2009.

[19] H. Schwender, K. Ickstadt, "Identification of SNP interactions using logic regression", *Biostatistics*, vol. 9, pp. 187-198, 2008.

[20] A. Katanforoush, M. Sadeghi, H. Pezeshk, E. Elahi, "Global haplotype partitioning for maximal associated SNP pairs", *BMC Bioinformatics*, vol. 10, 2009.

[21] X. Wan, C. Yang, Q. Yang, H. Xue, N. L. S. Tang, W. C. Yu, "Predictive rule inference for epistatic interaction detection in genome-wide association studies", *Bioinformatics*, vol. 26, pp. 30-37, 2010.

[22] Y. Liu, M. K. Ng, "Shrunken Methodology to Genome-wide SNPs Selection and Construction of SNPs Networks", *BMC Systems Biology*, vol. 4, 2010.

[23] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, "Gene ontology: tool for the unification of biology", *Nat Genet.*, vol. 25, pp. 25-29, 2000.

[24] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly and P. C. Sham, "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *The American Journal of Human Genetics*, vol. 81, pp. 559-575, Sep. 2007.

[25] Z. Li, S. Zhang, R. S. Wang, X. S. Zhang, L. Chen, "Quantitative function for community detection", *Phys. Rev. E*, vol. 77, pp. 036109, 2008.

[26] P. L. Wang, M. H. Dai, W. J. Xuan, R. C. McEachin, A. U. Jackson, L. J. Scott, B. Athey, S. J. Watson and F. Meng, "SNP Function Portal: a web database for exploring the function implication of SNP alleles," *Bioinformatics*, vol. 22, no. 14, pp. e523-e529, 2006.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE