Prioritization of Schizophrenia Risk Genes by a Network-Regularized Logistic Regression Method

Wen Zhang, Jhin-Rong Lin, Rubén Nogales-Cadenas, Quanwei Zhang, Ying Cai, and Zhengdong D. Zhang(✉)

Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA
zhengdong.zhang@einstein.yu.edu

Abstract. Schizophrenia (SCZ) is a severe mental disorder with a large genetic component. While recent large-scale microarray- and sequencing-based genome wide association studies have made significant progress toward finding SCZ risk variants and genes of subtle effect, the interactions among them were not considered in those studies. Using a protein-protein interaction network both in our regression model and to generate a SCZ gene subnetwork, we developed an analytical framework with Logit-Lapnet, the graphical Laplacian-regularized logistic regression, for whole exome sequencing (WES) data analysis to detect SCZ gene subnetworks. Using simulated data from sequencing-based association study, we compared the performances of Logit-Lapnet with other logistic regression (LR)-based models. We use Logit-Lapnet to prioritize genes according to their coefficients and select top-ranked genes as seeds to generate the gene sub-network that is associated to SCZ. The comparison demonstrated not only the applicability but also better performance of Logit-Lapnet to score disease risk genes using sequencing-based association data. We applied our method to SCZ whole exome sequencing data and selected top-ranked risk genes, the majority of which are either known SCZ genes or genes potentially associated with SCZ. We then used the seed genes to construct SCZ gene subnetworks. This result demonstrates that by ranking gene according to their disease contributions our method scores and thus prioritizes disease risk genes for further investigation. An implementation of our approach in MATLAB is freely available for download at: http://zdzlab.einstein.yu.edu/1/publications/LapNet-MATLAB.zip.

1 Introduction

SCZ is a common and severe lifelong brain disorder. It is a major cause of disability and reduces life expectancy by ~25 years on average. With its substantial mortality and morbidity, SCZ causes enormous personal and community burdens (Darves-Bornoz et al. 1995). In the United States, about 1 % of the general population, or 3 million

Electronic supplementary material The online version of this article (doi:10.1007/978-3-319-31744-1_39) contains supplementary material, which is available to authorized users.

F. Ortuño and I. Rojas (Eds.): IWBBIO 2016, LNBI 9656, pp. 434–445, 2016.
DOI: 10.1007/978-3-319-31744-1_39

Americans, suffer from this lifelong disabling illness (Regier et al. 1993). Thus, elucidating the etiology of the disease and developing effective treatment are of great medical urgency. The heritability of SCZ is well established. Recent studies have revealed a complex genetic architecture of the disease, involving multiple and heterogeneous genetic factors. Risk variants range in frequency from common to extremely rare and size from single nucleotide variants (SNVs) to large copy number variants (CNVs). Since 2009, GWASs have identified around 50 SCZ-associated loci with genome-wide statistical significance ($P < 5 \times 10^{-8}$) (Regier et al. 1993). Recently, a meta-analysis of SCZ discovered 108 risk loci, providing a significant source for identifying causal variants and causal genes of SCZ (Schizophrenia Working Group of the Psychiatric Genomics 2014). Rare congenital disorders associated with structural variants at 22q11.2, 15q13.3, 1q21.1, and several other genomic locations count for relatively small proportion of cases with SCZ (Bergen et al. 2012; Betcheva et al. 2013; Huang et al. 2010; Irish Schizophrenia Genomics and the Wellcome Trust Case Control 2012; Jeffrey A. Lieberman 2006; Shi et al. 2009; Shi et al. 2011; Wong et al. 2014). Increasing number of structural variation burden has been also observed in SCZ cases (Walsh et al. 2008).

Next-generation sequencing (NGS) has made it possible to evaluate the role of de novo or rare SNVs, both previously essentially inaccessible, in SCZ with DNA samples from parent-child trios or case-control cohorts. Using WES, instead of SNP microarray, as the genotyping tool to obtain a complete picture of genetic variants in coding sequences, a recent study assayed rare coding SNVs and small insertions and deletions (indels) in 2,536 SCZ cases and 2,543 normal controls and demonstrated a polygenic burden primarily arising from rare disruptive mutations distributed across many genes (Purcell et al. 2014). Recently, a number of statistical tests have been designed for WES-based variant analysis (Asimit and Zeggini 2010; Bansal et al. 2010; Basu and Pan 2011; Stitziel et al. 2011). Most of these methods first aggregate variants in each gene and then consider the association of each gene with the disease/phenotype separately. Hoffman et al. have developed a framework for applying a family of penalized regression methods that simultaneously consider multiple susceptibility loci in the same statistical model (Hoffman et al. 2013). In a more recent work, Larson and Schaid drew on penalized regression in combination with variant collapsing measures to identify rare variant enrichment in exome sequencing data (Larson and Schaid 2014).

Here we present a penalized regression method with graphical Laplacian network regularization and variant aggregation measures for case-control WES data analysis to assess gene contributions to the disease phenotypes. We first compared the performance of our regression method with other existing similar approaches using simulation under different scenarios. We then applied our method to the SCZ case-control WES data to prioritize SCZ risk genes. We discuss how the genes and pathways that we identified to make high contributions to SCZ may shed new light on genetic structure behind the SCZ in general.

2 Results and Discussion

Analysis of Simulated Phenotype and Genotype Data. We first simulated WES data sets with phenotypes under four different scenarios (Supplementary Table 1 and Supplementary Methods) and then used them to evaluate the performance of our network-regularized regression method (Logit-Lapnet) and three existing ones (Logit, Lasso, and Enet). Each simulation was replicated 50 times. After computing the sensitivity and specificity on the cutoff paths, we plotted the receiver-operating characteristic curve (ROC) and calculate the area under curve (AUC) of each method (Fig. 1). As its AUC is the largest under all four different simulation scenarios, the Logit-Lapnet method outperforms all other three. A similar performance assessment can also be made on the regularization path (Wan et al. 2013; Zhang et al. 2013) (Supplementary Fig. 1). From all the results, we could conclude that given available alternatives Logit-Lapnet is the best choice for prioritizing candidate genes among this class of algorithms.

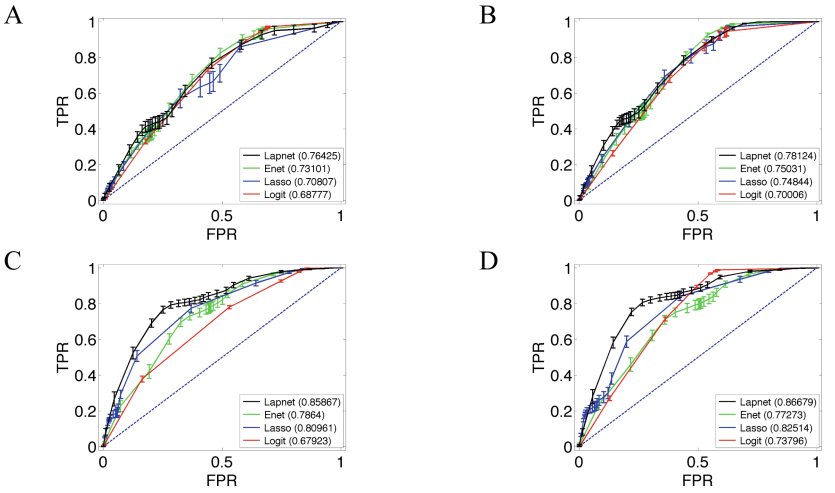

Fig. 1. ROCs with simulated samples. (A)–(D). Simulation under scenarios 1–4. All simulations were done with 100 samples. Average ROCs of four logistic regression methods are shown in different colors. The error bar indicates the standard deviation of all replicates for each sample. Their corresponding AUCs are given in parentheses in the figure legend.

To assess how the sample size affects the performance, we also simulated data sets with different sample sizes ranging from 10 to 200 under each scenario. For each sample size, we simulated every model 40 times and calculated the average AUCs (Supplementary Figs. 2 and 3). Over all, the LR and its extensions gave good analysis results for simulated case-control phenotype-genotype data– the AUC ranged from 0.6 to 0.9. The assessment of the Lasso and the Enet methods is less clear: one outperforms the other in each of two simulation scenarios. We also calculated and compare the F1

scores of regression methods. F1 score is the harmonic mean of precisions and recalls and acts as the integration of both of these two evaluations. This property makes it an informative and efficient measurement of performance of different methods. Here, the plot shows that under all four simulation conditions the F1 scores of the Logit-Lapnet method are the highest among four methods being studied (Supplementary Fig. 4). The F1 score comparison indicates that the Logit-Lapnet algorithm is more accurate than Logit, Lasso, and Enet methods.

Due to feasibility and efficiency of the Logit-Lapnet algorithm in prioritizing risk genes, we could further apply the method to real WES data set and get prioritizations of the genes so as to identify important genes relating to the disease that under consideration. Our motivation to develop the method for association studies is based on the hypothesis that integration of interaction networks improves prediction precision of logistic models. The network-constrained algorithm has been proved to out-perform alternative options such as lasso and elastic net analyses that are implemented separately from biological input (Li and Li 2008; Wan et al. 2013; Zhang et al. 2013). Enlightened by the application of Logit-Lapnet to efficiently identify molecular pathways and cancer biomarkers, we adapted this class of methods to analyze a set of WES data for SCZ. We use the ratio matrix of damaging allele counts over neutral allele counts to represent normalized population genotype information. The Logit-Lapnet approach is more sensitive for identifying disease genes because the relevant network modules are considered by using the regularization based on the network. Laplacian graphs are derived from gene networks, for which we used High-quality INTeractomes (HINT) network in our study. HINT is a database of high-quality protein-protein interactions in different organisms, which have been compiled from different data sources and then filtered both systematically and manually to remove erroneous and low-quality interactions (Das and Yu 2012).

Combining information of gene interactions, the Laplacian graphs form the penalized term with regard to contribution coefficient of each gene. The L_2-normalized item incorporates network information into the estimation procedure of the regression model and encourages smoothness in the estimate of contributions of candidate genes. Incorporation of a gene network contributes to the advantages of Logit-Lapnet over the other methods since in this way the method integrates into its calculation a vast amount of *a priori* biological information from the network, which is ignored in either lasso or elastic net methods. In summary, our method takes advantage of the information obtained about genotype relationships beyond the scope of other single regression study.

Analysis of SCZ WES Data. The simulation results clearly indicate that the LR and its extensions can be effectively applied to case-control genotype data to identify genes related to the phenotype or disease under consideration and the Logit-Lapnet method gives the best performance. Here, we applied this method to the SCZ WES data to estimate the corresponding coefficients as phenotypic contributions of SCZ target genes. First, we derived from the WES data the phenotype vector $\mathbf{y}$, the gene evaluation data matrix $\mathbf{X}$, and the normalized graph Laplacian matrix $\mathbf{L}$. Corresponding to the SCZ patient cohort, $\mathbf{y}$ is a binary column vector with the elements: 1's for patients with SCZ and 0's for ones without SCZ. $\mathbf{X}$ and $\mathbf{L}$ are $n \times 844$ and 844×844 matrices,

respectively, where n is the number of individuals (cases and controls) and 844 is the number of candidate genes, some of which could be responsible for SCZ in the original WES study cohort. In the SCZ WES data set from dbGaP, there are 2,545 cases and 2,545 controls in total. Maximizing the Logit-Lapnet function (Eq. 3) is a computationally intensive process (Supplementary Fig. 5), which necessitates parallel processing of large data sets. We randomly divided the SCZ WES data set into 25 subsets, each with $\sim$100 cases and $\sim$100 controls. We analyzed them in parallel and then integrated gene scores. Using the Logit-Lapnet method, we estimated from these subsets in parallel the coefficients of candidate genes as their contributions to SCZ. We arranged the genes in each list in descending order of their coefficients and integrated the ranked gene lists using a robust rank aggregation method (Kolde et al. 2012). We randomized the SCZ WES data set to evaluate this ranked gene list and to remove possible false positives. In each iteration, we randomized the disease labels among the samples and processed the random data set in the same manner as the real one. After many iterations of randomization, for each gene we calculated the probability that the rank of this gene based on a random data set is the same as or even better than that based on the real data set. After removing genes with probability ≥0.05, we then considered the top 20 genes as the most promising candidates for SCZ in the WES study cohort (Supplementary Table 2). 10 of them including CARD10, TIMP2, PPP2CA, and PTPRB were also identified as SCZ risk genes by the original exome sequencing study (Purcell et al. 2014).

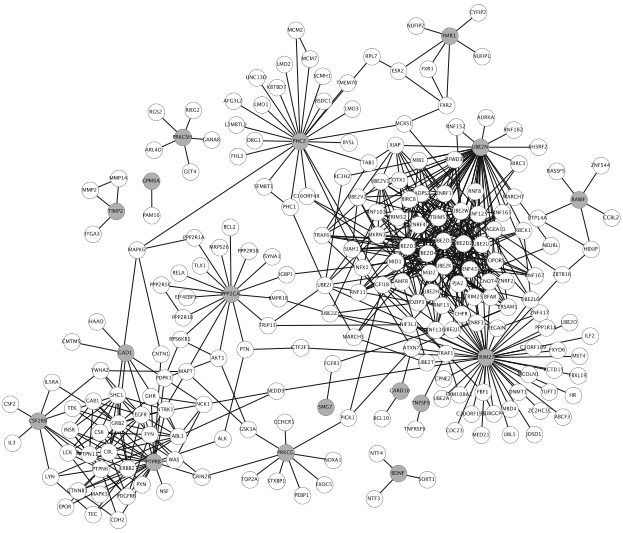

Fig. 2. SCZ gene subnetwork. After filtering by randomization, 20 genes with largest regression coefficients were used as seeds to form the subnetwork in HINT. Gray and white nodes represent the seed genes and their direct neighbors, respectively, in the network. Three seed genes are not included in HINT and thus are omitted from the subnetwork.

Because most cellular components exert their functions through interactions with other cellular components, such inter- and intracellular inter-connectivity implies that the impact of a specific genetic variation is not restricted to the activity of an SCZ-related gene product that carries it, but can spread along the links of the network and alter the activity of other SCZ-related gene products that otherwise carry no changes. Therefore, an understanding of SCZ genes' network context is essential to understand the genetics of this disease. Using the aforementioned top 20 genes as seeds and the 'extraction by shell' method implemented in SubNet (Lemetre et al. 2013), we extracted a GsN (Fig. 2) from HINT (Das and Yu 2012), a high-quality protein-protein interaction network. This SCZ GsN contains 223 proteins and 546 interactions among them. The majority of the proteins (207, 92.83 %) form a connected component. Despite its small size, this subnetwork clearly shows a power law distribution for its node degrees: a prominent characteristic of complex biological networks. Unlike the majority of proteins in the subnetwork, a few of them have a large number of interaction partners and are functionally more important. Many seed genes are such network hubs in the SCZ GsN. PHC2, one network hub, was found to be affected by mutations in SCZ patients (Purcell et al. 2014). Known SCZ loci 5q31.1, 6p22, and 12q22 contain, respectively, PPP2CA, TRIM27, and UBE2 N, which are among top 20 risk genes that we identified for SCZ in the WES study cohort. In addition, these three genes contain nonsynonymous coding variants with minor allele frequencies (MAF) less than 0.1 %. Located in 12q15-q21, PTPRB, another seed gene with high contribution, contain nonsynonymous coding variants with MAF < 0.5 %.

As a proof of principle, our SCZ study demonstrates that our data analysis workflow and methods can be successfully applied to WES data sets to identify disease risk genes and subnetworks. Although they are designed to be applicable to large GWAS and WES data sets, such as those provided by dbGaP, the implementation in main text of this paper processes data sets with moderate sample sizes. Because individual genotypes have larger effect on transcript abundance than on disease risk, a small sample size can still be powerful to detect disease variants and genes (Gibson 2014). To analyze larger ones, the optimization problem – the rate limiting step – needs to be solved by more efficient optimizers or parallel computing or both. This difficulty is still an open challenge and an active research area. For comparison of running time with larger different sample sizes, please see Supplementary Fig. 5.

3 Methods

Input data for WES Regression Analysis. After SNVs and indels in the sequenced subjects are identified and their genotypes called, while assuming the reference alleles (RAs) of these variants to be neutral, we predicted the functional consequence (i.e., neutral/tolerable/benign or damaging/deleterious) of the alternative alleles (AAs) using computational programs. Let n and p be the numbers of genes and sequenced subjects (samples), respectively. To carry out the regression analysis of the case-control WES data, we first summarized the genotypes and the allelic functional annotations on gene level in two n-by-p matrices, **D** and **N**, which hold all of the numbers regarding the

neutral and damaging allele counts of each gene i respectively in each sample j. As longer genes tend to have more variants, to prevent the gene length from skewing up the analysis, we normalized the damaging allele counts through the way of getting them divided by neutral allele counts. Since each gene may contain multiple variants (SNPs and indels), we counted the damaging and the neutral alleles of all variants within a gene. In this way, multiple variants mapped to one gene are combined to obtain the allele counts on the gene level. Most of the sequencing errors are filtered out at the quality control steps as part of the variant calling process. For any remaining ones, because they occur randomly, their effect will be cancelled out in a case-control study. Thus, we define the input ratio matrix $\mathbf{X}$ as in Eq. 1.

$$X := [X_{ij}], D := [D_{ij}], N := [N_{ij}], X_{ij} = \frac{D_{ij}}{N_{ij} + \tau}, \tau = \begin{cases} 0 & \text{if } N_{ij} \neq 0 \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

Due to computational intensity, instead of scoring the whole gene set, the current implementation of our method scores and thus prioritizes a set of genes preselected based on prior knowledge given the genotypic data. This set of genes can be obtained in two steps. First, a core group of genes can be collected from various disease gene databases (Goh et al. 2007; Pinero et al. 2015; Pletscher-Frankild et al. 2015; Rappaport et al. 2014). Then, adding their neighboring genes in a gene network can augment this core group of genes. Even for less commonly studied diseases, this approach can procure a set of relevant genes for scoring. Genes selected for scoring and prioritization are included in matrices $\mathbf{X}$ and $\mathbf{L}$ for analysis. It is the disease, not the sample of the data set, that determines what genes to be selected. Therefore, the same set of genes included in $\mathbf{X}$ and $\mathbf{L}$ will be used for two different samples of the same disease. If the disease has different genetic risk factors in these two samples, then the ranking of selected genes will be different for these two samples in the results. For flowchart to further illustrate the process of input data integration, please refer to Supplementary Fig. 6. Pre-processing of the WES data by generating the ratio data matrix is a novelty of the method. Current studies do not provide this kind of data matrix generation in analyzing WES data.

Graphically Laplacian Network-Regularized LR Method. In a recent study a network-regularized linear regression method has been proposed for variable selections (Li and Li 2008). Compared with Lasso and elastic net (Enet) methods, they show advantages of using the network-regularized process, which include higher precisions and comparable or even better sensitivity and specificity. In other variable selection problems where the output vectors only contain binary values (i.e., either 0 or 1), the LR with network regularization outperforms previous alternatives (Zhang et al. 2013). The Laplacian graphical network-regularized LR methods (Logit-Lapnet) have been used to identify molecular pathways of breast cancers (Zhang et al. 2013). Its application to simulated gene expression data showed excellent sensitivity and specificity and higher accuracy than Lasso and Enet methods. Inspired by its original application, we adapted the Logit-Lapnet method with significant redesign to analyze the case-control genotype data, focusing on NGS-generated data. Different from previous

studies which use Logit-Lapnet to analyze the gene expression data (Zhang et al. 2013), we utilize the method to WES data analysis.

Let $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ be the phenotype vector, where $y_i = 0$ for control and 1 for case ($i = 1, \ldots., n$). Given matrix $\mathbf{X}$, the aforementioned gene evaluation data matrix $\mathbf{Y}$ is modeled by logistic function:

$$Y = \Pr(y = 1|\mathbf{X}; \boldsymbol{\theta}) = \frac{e^{\mathbf{X}\theta}}{1 + e^{\mathbf{X}\theta}} \tag{2}$$

The regression coefficients in $\boldsymbol{\theta}$ quantify the effect of the genotypes of genes on the 'odds ratio' of having the disease and thus represent the importance of genes. Genes with larger coefficients will be ranked higher than others (Li and Li, 2008). The coefficient vector $\boldsymbol{\theta}$ is estimated by minimizing the negative log-likelihood function, or equivalently, maximizing the positive function, of logistic model combined with penalized terms. The mitigated formula of logistic graph Laplacian net criteria is

$$\mathbb{C}(\boldsymbol{\theta}, \lambda, \alpha) = \sum_{i=1}^{n} \left[-y_i X_i \boldsymbol{\theta} + \ln\left(1 + e^{X_i \boldsymbol{\theta}}\right)\right] + \lambda\alpha|\boldsymbol{\theta}|_1 + \lambda(1 - \alpha)\boldsymbol{\theta}^T \mathbf{L}\theta \tag{3}$$

where X_i is the i-th row-vector of data matrix $\mathbf{X}$. $\mathbf{L}$ is the normalized graph Laplacian matrix, $|\boldsymbol{\theta}|_1$ is the L_1 norm of $\boldsymbol{\theta}$, i.e., $|\boldsymbol{\theta}|_1 = \sum_{j=1}^{p} \theta_j$, with θ_j corresponding contribution coefficient of each gene. Suppose $\mathbf{A}$ and $\mathbf{E}$ are adjacency matrix and degree matrix of the network, respectively, then $\mathbf{L}$ is given as:

$$\mathbf{L} = \mathbf{I} - \mathbf{E}^{-\frac{1}{2}} \mathbf{A} \mathbf{E}^{-\frac{1}{2}} \tag{4}$$

where $\mathbf{I}$ is the identity matrix with the same dimension to that of $\mathbf{A}$ or $\mathbf{E}$.

Equation 3 contains three terms: the negative log-likelihood function; the L1 normalized penalty term, which L1 penalizes the norm of θ; and the graph Laplacian term, which is formulated as the inner product of θ regarding to Laplacian matrix L in (Zhang et al. 2013). The last term can be treated as the L2 normalized item as well. Equation 3 also makes it clear that the Laplacian network-regularized LR is the ordinary form of general logistic regressions, which include several more special types of logistic models. In case when L = I, the algorithm becomes an Enet module. When $\alpha = 1$ and $\lambda \neq 0$, the method is regressed to Lasso. If $\lambda = 0$, the method is further simplified as a standard LR model without penalties (referred to as the Logit model hereafter). It is clear that Logit-Lapnet, Enet, Lasso, and Logit are LR methods with different levels of constraints on regression coefficients, known as contributions in our methods. Given the data matrix X, the optimal values for the model parameters α and λ are determined by a leave-one-out cross validation (CV) procedure, and the optimal coefficients θ are estimated by minimizing the criteria $\mathbb{C}(\theta, \lambda, \alpha)$ given optimal parameters λ_{opt} and α_{opt}:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathbb{C}\left(\theta, \lambda_{\text{opt}}, \alpha_{\text{opt}}\right) \tag{5}$$

Convexity of Eq. 3 guarantees minimal index performance $\mathbb{C}\left(\theta^*, \lambda_{\text{opt}}, \alpha_{\text{opt}}\right)$, and the corresponding θ^* could be worked out by a standard optimizer such as the Matlab-based software CVX (Grant and Boyd 2014). For theoretical properties of Logit-Lapnet, please refer to Lemma 1, 2, 3 and Theorem 1 in (Zhang et al. 2013) and the references therein. The Lemma 3 and Theorem 1 in (Zhang et al. 2013) provide grouping effects for the Logit-Lapnet procedure. Mathematical formulation in (Zhang et al. 2013) supports our novel application of Logit-Lapnet to risk gene prioritization.

Processing SCZ WES Genotype Data. The Swedish SCZ population-based case and control WES data set (study accession ID number: phs000473.v1.p1) was downloaded from dbGaP. Excluding the variants in non-coding regions, we extracted two sets of variants – SNVs and indels – in exons of the sequenced subjects from the WES data set. For SNVs, the functional roles of their AAs were predicted by SIFT (Hu and Ng 2013), PolyPhen2 (Sachdev and Keshavan 2010), and Blosum62 scoring matrices. We combined these three scores using a so-called 'damaging-dominant' rule – the AA was considered as damaging as long as one method predicts so. We used this policy since the variant annotations are usually predicted by and selected among high impacts and damaging, in this sense, is the higher impact compared with neutral/tolerant. For indels, we used SIFT to predict the functional roles of their AAs. RAs were considered neutral in this regard. We derived the **D** and **N** matrices (Eq. 1) for SNVs and indels separately from their functional annotations. After combining **D** and **N** of SNVs and indels separately, we calculated **X** and use it as the input data matrix. The ratio matrix reflects the relative genetic influences of damaging alleles in each gene on the disease status.

Compiling SCZ-related Genes. Derived from the SCZ WES genotype data, the input matrix X holds the damage load for 13,899 sequenced genes in 200 samples. Given the large number of genes and the modest sample size, to keep the statistical analysis tractable we focused our analysis on genes likely to be related to SCZ according to prior knowledge. Our strategy for gene selection was to include both genes most relevant to SCZ and ones with potential but unknown associations with SCZ. We used a two-tier approach. First, we compiled a list of 308 genes that have been shown to be SCZ-related:

- 217 genes with prior evidence for association with SCZ, which are prioritized in the data source SZGR (Jia et al. 2010).
- 91 genes collected from published literatures (Supplementary Table 3).

Next, based on the 'guilt-by-association' principle, we collected 536 direct neighbors of these 308 genes in the HINT. Together, we selected 844 candidate genes.

Networks used in SCZ WES Data Analysis. We used HINT with the Logit-Lapnet method for our SCZ WES data analysis. From HINT, the Laplacian matrix of the aforementioned target genes was generated and then used as the graphical Laplacian normalized term in Eq. 2.

4 Conclusion

We developed a computational framework for WES data analysis that combines both prioritization of disease risk genes by graphical Laplacian regularized LR and extraction of disease-related GsN by SubNet. Although the Logit-Lapnet method has been used before to analyze gene expression data and somatic mutation profiles (Betcheva et al. 2013; Hoffman et al. 2013; Shi et al. 2009; Stitziel et al. 2011), our study demonstrates here that after data transformation it can also be efficiently applied to exome sequencing-based GWAS genotype data. Method assessment by simulation shows that Logit-Lapnet is more sensitive for identifying seed genes with higher priorities than other related methods. We applied our method to SCZ WES data. Top-ranked genes are either known SCZ risk genes or closely related to SCZ. Using them as seeds, we extracted the SCZ GsN from known protein interaction network. It provides a valuable subnetwork for pathway and gene module detection of SCZ.

Acknowledgements. This work was supported by the NIH Pathway to Independence Award from National Library of Medicine (5R00LM009770-06) and the American Heart Association Grant-in-Aid (13GRNT16850016) to Z.D.Z.

References

Asimit, J., Zeggini, E.: Rare variant association analysis methods for complex traits. Ann. Rev. Genet. **44**, 293–308 (2010)

Bansal, V., Libiger, O., Torkamani, A., Schork, N.J.: Statistical analysis strategies for association studies involving rare variants. Nature Rev. Genet. **11**, 773–785 (2010)

Basu, S., Pan, W.: Comparison of statistical tests for disease association with rare variants. Genet. Epidemiol. **35**, 606–619 (2011)

Bergen, S.E., O'Dushlaine, C.T., Ripke, S., Lee, P.H., Ruderfer, D.M., Akterin, S., Moran, J.L., Chambert, K.D., Handsaker, R.E., Backlund, L., et al.: Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. Mol. Psychiatry **17**, 880–886 (2012)

Betcheva, E.T., Yosifova, A.G., Mushiroda, T., Kubo, M., Takahashi, A., Karachanak, S.K., Zaharieva, I.T., Hadjidekova, S.P., Dimova, I.I., Vazharova, R.V., et al.: Whole-genome-wide association study in the Bulgarian population reveals HHAT as schizophrenia susceptibility gene. Psychiatr. Genet. **23**, 11–19 (2013)

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.: The variant call format and VCFtools. Bioinformatics **27**, 2156–2158 (2011)

Darves-Bornoz, J.M., Lemperiere, T., Degiovanni, A., Gaillard, P.: Sexual victimization in women with schizophrenia and bipolar disorder. Soc. Psychiatry Psychiat. Epidemiol. **30**, 78–84 (1995)

Das, J., Yu, H.: HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst. Biol. **6**, 92 (2012)

Gibson, G.: A primer of human genetics (Sinauer Associates, Inc.) (2014)

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabasi, A.L.: The human disease network. Proc. National Acad. Sci. US Am. **104**, 8685–8690 (2007)

Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 2.1 (2014)

Hoffman, G.E., Logsdon, B.A., Mezey, J.G.: PUMA: A unified framework for penalized multiple regression analysis of GWAS data. PLoS Comput. Biol. **9**, e1003101 (2013)

Hu, J., Ng, P.C.: SIFT Indel: Predictions for the functional effects of amino acid insertions/deletions in proteins. PLoS ONE **8**, e77940 (2013)

Huang, J., Perlis, R.H., Lee, P.H., Rush, A.J., Fava, M., Sachs, G.S., Lieberman, J., Hamilton, S.P., Sullivan, P., Sklar, P., et al.: Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression. Am. J. Psychiatry **167**, 1254–1263 (2010)

Irish Schizophrenia Genomics, C., and the Wellcome Trust Case Control, C. Genome-wide association study implicates HLA-C*01:02 as a risk factor at the major histocompatibility complex locus in schizophrenia. Biological psychiatry **72**, 620–628 (2012)

Jeffrey, A., Lieberman, T.S.S., Perkins, D.O.: Textbook of Schizophrenia, The American Psychiatric Publishing, Arlington, Virginia, USA (2006)

Jia, P., Sun, J., Guo, A.Y., Zhao, Z.: SZGR: A comprehensive schizophrenia gene resource. Mol. Psychiatry **15**, 453–462 (2010)

Kim, S., Jeong, K., Bafna, V.: Wessim: A whole-exome sequencing simulator based on in silico exome capture. Bioinformatics **29**, 1076–1077 (2013)

Kolde, R., Laur, S., Adler, P., Vilo, J.: Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics **28**, 573–580 (2012)

Larson, N.B., Schaid, D.J.: Regularized rare variant enrichment analysis for case-control exome sequencing data. Genet. Epidemiol. **38**, 104–113 (2014)

Lemetre, C., Zhang, Q., Zhang, Z.D.: SubNet: A Java application for subnetwork extraction. Bioinformatics **29**, 2509–2511 (2013)

Li, C., Li, H.: Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics **24**, 1175–1182 (2008)

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al.: Targeted capture and massively parallel sequencing of 12 human exomes. Nature **461**, 272–276 (2009)

Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., Furlong, L.I.: DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. Database J. Biol. Databases Curation **2015**, 28 (2015)

Pletscher-Frankild, S., Palleja, A., Tsafou, K., Binder, J.X., Jensen, L.J.: DISEASES: Text mining and data integration of disease-gene associations. Methods **74**, 83–89 (2015)

Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kahler, A., et al.: A polygenic burden of rare disruptive mutations in schizophrenia. Nature **506**, 185–190 (2014)

Rappaport, N., Twik, M., Nativ, N., Stelzer, G., Bahir, I., Stein, T.I., Safran, M., Lancet, D.: Malacards: A comprehensive automatically-mined database of human diseases. Current protocols in bioinformatics/editoral board, Andreas D Baxevanis [et al.] 47, 1 24 21–21 24 19 (2014)

Regier, D.A., Narrow, W.E., Rae, D.S., Manderscheid, R.W., Locke, B.Z., Goodwin, F.K.: The de facto US mental and addictive disorders service system. Epidemiologic catchment area prospective 1-year prevalence rates of disorders and services. Arch. Gen. Psychiatry **50**, 85–94 (1993)

Sachdev, P.S., Keshavan, M.S.: Secondary Schizophrenia. United Kingdom at the University Press, Cambridge (2010)

Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L., Nolan, G.P.: Computational solutions to large-scale data management and analysis. Nature Rev. Genet. **11**, 647–657 (2010)

Schizophrenia Working Group of the Psychiatric Genomics Consortium: Biological insights from 108 schizophrenia-associated genetic loci. Nature **511**, 421–427 (2014)

Shi, J., Levinson, D.F., Duan, J., Sanders, A.R., Zheng, Y., Pe'er, I., Dudbridge, F., Holmans, P. A., Whittemore, A.S., Mowry, B.J., et al.: Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature **460**, 753–757 (2009)

Shi, Y., Li, Z., Xu, Q., Wang, T., Li, T., Shen, J., Zhang, F., Chen, J., Zhou, G., Ji, W., et al.: Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. Nat. Genet. **43**, 1224–1227 (2011)

Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., Ng, P.C.: SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. **40**, W452–W457 (2012)

Stitziel, N.O., Kiezun, A., Sunyaev, S.: Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol. **12**, 227 (2011)

Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., Nord, A.S., Kusenda, M., Malhotra, D., Bhandari, A., et al.: Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science **320**, 539–543 (2008)

Wan, Y.W., Nagorski, J., Allen, G.I., Li, Z.H., Liu, Z.D.: Identifying cancer biomarkers through a network regularized Cox model. Genomic Signal Processing and Statistics (GENSIPS), 2013 IEEE International Workshop on (Houston, pp. 36–39. IEEE, TX (2013)

Wang, K., Li, M., Hakonarson, H.: ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. **38**, e164 (2010)

Wong, E.H., So, H.C., Li, M., Wang, Q., Butler, A.W., Paul, B., Wu, H.M., Hui, T.C., Choi, S.C., So, M.T., et al.: Common variants on Xq28 conferring risk of schizophrenia in Han Chinese. Schizophr. Bull. **40**, 777–786 (2014)

Zhang, W., Wan, Y.W., Allen, G.I., Pang, K., Anderson, M.L., Liu, Z.: Molecular pathway identification using biological network-regularized logistic models. BMC Genom. **14**(Suppl 8), S7 (2013)