

·信息研究·

基于知识融合策略构建双相障碍致病基因网络

刘 轲¹, 赵 虎¹, 刘 燕¹, 饶绍奇^{2,3*}

(1.中山大学中山医学院,广东 广州 510080;2.广东医学院公共卫生学院,广东 东莞 523808;3.中山大学公共卫生学院,广东 广州 510080)

摘 要:【目的】提出基于知识融合策略构建基因网络方法,并应用于双相障碍相关的致病基因网络分析。【方法】将 Wellcome Trust Case Control Consortium (WTCCC)提供的双相障碍全基因组单核苷酸多态(SNP)数据与人类蛋白质-蛋白质互作数据库对应的基因做交集。通过单体型全模型 logistic 回归模型检验获得经多重检验校正统计学显著的基因互作对子,并由此构建致病基因网络以及挖掘连通度显著高于理论分布的核心致病基因。【结果】采用知识融合的方法,将数据维度从 482 248 个 SNP 位点降至 98 157。经统计模型检验获得 3 841 个互作基因用于构建双相障碍致病基因网络,并挖掘出 115 个核心致病基因。其中,在连通度高于 30 的 29 个核心基因中,有 12 个重复了以前的报道(PRKCA, EGFR, ESR1, ATXN1, FYN, CREBBP, TP53, AKT1, CSNK2A1, DLG1, PTN 和 LYN),另外 17 个未被报道过的基因从其生物功能以及致病分子机制上看,可能是新的双相障碍易感基因(SMAD3, SRC, GRB2, PIK3R1, ZBTB16, ABL1, APP, EP300, TGFBR1, SYK, YWHAZ, INSR, MAPK1, PRKCB, PRKCD, SMAD2 和 SVIL)。【结论】本文提出的基于蛋白质-蛋白质互作知识引导的基因网络构建方法是一种可靠的系统性分析方法,有助于全面地了解复杂疾病的分子网络机制和确立核心风险基因。

关键词: 双相障碍;知识学习;蛋白质-蛋白质互作;全基因组关联;基因网络;核心风险基因

中图分类号:R739

文献标志码:A

文章编号:1672-3554(2013)03-0471-07

Construction of Responsible Gene Network for Bipolar Disorder by Using a Knowledge-integration-based Strategy

LIU Ke¹, ZHAO Hu¹, LIU Yan¹, RAO Shao-qi^{2,3*}

(1. Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou 510080, China; 2. School of Public Health, Guangdong Medical College, Dongguan 523808, China; 3. School of Public Health, Sun Yat-Sen University, Guangzhou 510080, China)

Abstract: 【Objective】 To propose a knowledge-fusion strategy for constructing gene networks, and to apply the proposed approach to analyze the gene networks for bipolar disorder (BPD). 【Methods】 The intersecting gene set between all genes in a protein-protein interaction (PPI) database and the genes in the whole genome single nucleotide polymorphism (SNP) dataset for BPD, provided by the Wellcome Trust Case Control Consortium (WTCCC), were defined. Statistically significant epistatic gene pairs in the gene set were then obtained by using a haplotype-based full-model logistic regression and multiple testing corrections. Finally, a disease-causing gene network was constructed by using these epistatic gene pairs, and the hub genes whose connection degrees were significantly higher than theoretical ones were identified. 【Results】 By using the proposed knowledge-fusion strategy, the data dimension for the whole genome SNP data for BPD was reduced, from 482,248 SNPs to 98,157 SNPs. A total of 3,841 genes with significant epistasis identified by using the abovementioned statistical model were used to construct the underlying gene network for BPD, from which 115 hub-genes were found. Among 29 hub-genes with degrees more than 30, 12 were found repeating the previous findings (PRKCA, EGFR, ESR1, ATXN1, FYN, CREBBP, TP53, AKT1, CSNK2A1, DLG1, PTN, and LYN), while the remaining 17 genes might be novel susceptibility genes, judged by their functional involvements (SMAD3, SRC, GRB2, PIK3R1,

收稿日期:2012-11-14

基金项目:国家自然科学基金(30830104,31071166);广东省自然科学基金(8251008901000007);广东省科技计划攻关项目(2009A030301004);东莞市科技重点项目(201108101015);广东医学院基金项目(XG1001,XZ1105,STIF201122,JB1214)

作者简介:刘轲,硕士生,研究方向:生物信息学,E-mail: liukebruce@163.com; *通信作者:饶绍奇,博士,教授,E-mail: raoshaoq@gdmc.edu.cn

ZBTB16, ABL1, APP, EP300, TGFBR1, SYK, YWHAZ, INSR, MAPK1, PRKCB, PRKCD, SMAD2, and SVIL). 【Conclusion】 This real data analysis demonstrates that the proposed gene network approach, which was guided by PPI knowledge, was a reliable systematic method, and would help us have a global view on the underlying molecular networking mechanisms for complex diseases, and would help us find the hub risk genes.

Key words: bipolar disorder; knowledge fusion; protein-protein interaction; genome-wide association; gene network; hub genes

[J SUN Yat-sen Univ(Med Sci), 2013, 34(3):471-477]

双相障碍(bipolar disorder, BPD), 又称为躁狂抑郁症, 临床症状主要表现为躁狂状态与抑郁状态交替出现, 大多数病人有反复发作的倾向, 终身患病风险估计值高达 4%^[1]。BPD 可导致患者的社会功能损害与自杀, 对社会造成严重危害^[2]。BPD 确切的病因尚不清楚, 目前所公认的是多种致病因素(生物-心理-社会)所导致。迄今对该病病因学的研究探索主要聚焦于神经生化、神经内分泌、神经影像学、心理社会因素。遗传学研究表明 BPD 属于一种多基因及环境共同作用的复杂遗传疾病, 遗传因素能解释高达 60% ~ 85% 的发病风险^[3-4]。在过去数十年中, 传统基于家系的连锁分析和基于人群的病例对照候选基因关联分析一直是识别 BPD 易感基因的常用手段^[5]。随着第三代多态性遗传标记单核苷酸多态性(single nucleotide polymorphism, SNP)的出现, 基于全基因组 SNP 数据的关联分析也被应用于大规模扫描 BPD 新的易感基因位点^[6]。目前的三种分析策略多采用的是单位点、单基因分析方法, 尚缺乏针对 BPD 的复杂分子互作机制及基因网络系统性地分析。构建 BPD 基因互作网络, 能够更深入地探索各基因的协同机制对 BPD 发生发展的作用。目前已发展了多种基于数学算法和模型的基因网络构建方法, 例如用来表现同一层面上基因的共同特性, 如相似的功能或调控原理的聚类分析技术^[7]; 或是通过逻辑规则(布尔函数)从基因的表达状态推断基因之间互作关系的布尔网络模型^[8]; 或是运用贝叶斯概率模型表示多个变量的随机多元统计关系的贝叶斯网络^[9]。但是这些基于数学模型的方法都存在一些不足。首先, 这些数学方法忽略了基因网络的生物学特性, 单纯地从数学上推理, 结果可释性较差。其次, 针对高通量组学数据, 计算量大和结果稳定性差。为了克服传统网络构建方法应用于高通量 SNP 数据可能带来的一些问题, 本研究提出了一种基于知识学习的基因网络构建方法, 采用以人类蛋白质-蛋白质互作

(protein-protein interaction, PPI) 知识和数据库为引导、提取候选基因互作对子的网络构建新策略。其基本思想是在确定基因产物间存在某种生物学互作(如 PPI)的前提下, 逆向通过检测疾病特异的基因对子间 SNP 的统计互作关系, 构建与疾病相关的基因网络, 或称致病基因网络。通过进一步网络分析, 确定导致疾病的核心基因, 即维持生命基本活动和人类健康的必需基因。

1 材料和方法

1.1 数据来源

申请下载 Wellcome Trust Case Control Consortium(WTCCC) 提供的全基因组 SNP 组学数据^[10]。数据由该机构的 50 个团队历时 50 年收集英国人群所得, 包括 7 种人类复杂遗传疾病, 14 000 疾病样本和 3 000 健康对照样本。本研究采用了其中的 BPD 部分。

1.2 数据预处理

WTCCC 提供的 BPD 数据包括了 1 868 个 BPD 患者和 2 938 个健康对照, 共计 482 247 个 SNP 的数据。为了控制数据质量, 剔除不满足下列标准之一的 SNP 数据: Hardy-Weinberg 平衡定律(经 Bonferroni 校正 $P < 5 \times 10^{-7}$)、数据缺失率超过 5%、最小等位基因频率(Minor Allele Frequency, MAF) 小于 5%。并以连锁不平衡(Linkage disequilibrium, LD)量度 $R^2 > 0.8$ 的标准遴选标签 SNP。所有的数据处理由 Plink 软件计算分析完成。

1.3 SNP 映射到基因

通过 NCBI 数据库下载了目前已经实验验证的全部基因信息以及 SNP 注释信息。依托 R 语言程序平台, 编写用于提取 SNP 相对位置及其对应基因的程序。如果 SNP 未落在任一基因内, 采用搜索该 SNP 临近 ± 5 kb 基因的方法定位。删除所有经过上述操作仍然无法映射的 SNP 位点。

1.4 数据降维

为了避免传统的单位点分析降维方法错失重要的微效基因的缺陷,采用基于数据库的生物学互作知识融合的策略。具体步骤如下:①下载人类蛋白质参考数据库(Human Protein Reference Database, HPRD)中的 PPI 数据库部分。截止到 2012 年 9 月 21 日,已有 39 194 对蛋白-蛋白互作数据被收录。②提取 SNP 映射基因与 PPI 数据库里基因的交集,获得用于 SNP-SNP 互作分析的候选基因集合。

1.5 致病基因网络的构建

以两基因间最显著的互作 SNP 结果为依据,采用基于单体型的全模型 logistic 回归检验^[11]的方法确定 SNP 位点间的互作:

$$\log \frac{P(D = 1|x_1, x_2)}{P(D = 0|x_1, x_2)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \quad (1)$$

上述方程式(1)中, $D = 1$ 表示病例人群, $D = 0$ 表示对照人群; x_1, x_2 分别表示在单体型构型中两个 SNP 的等位基因信息, $x_1 x_2$ 表示交互项, β_1, β_2 分别代表两个位点的主效应回归系数。通过检测交互项系数 β_{12} 是否为零判定两 SNP 位点间是否存在交互效应。假设有两个 SNP 位点 G 和 H,其等位基因分别为 $\{G_1, G_2\}$ 和 $\{H_1, H_2\}$,其中 G_1 和 H_1 为两个低频等位基因,依单体型构型回归模型中自变量的编码如表 1。

表 1 基于单体型 logistic 回归模型自变量编码
Table 1 The codes of independent variables in the haplotype-based logistic regression model

	x_1	x_2	$x_1 x_2$
$G_1 H_1$	1	1	1
$G_1 H_2$	1	0	0
$G_2 H_1$	0	1	0
$G_2 H_2$	0	0	0

单体型频率通过期望最大化算法(expectation maximization algorithm, 简称 EM 算法)估算^[12]。由于两基因互作检验涉及多个 SNP 对子的检验,采用 Bonferroni 法校正多重检验显著性水平。以上全部运算过程采用 Windows 系统平台的 R 语言编写程序完成。最后将具有显著统计互作关系的基因两两连接,建立等权重无向基因网络。基因网络图由 Cytoscape 软件绘制。

1.6 网络拓扑属性分析

致病基因网络的拓扑属性分析包括以下几个参数^[13]:

①连通度(degree),指网络中直接与某一节点相连的边的数目。

②聚类系数(clustering coefficient, CC),反映相邻两个基因间共同调控基因的重合度,定义为:

$$CC_v = \frac{n}{C_k^2} = \frac{2n}{k(k-1)} \quad (2)$$

方程式(2)中 n 表示节点 v 的所有 k 个邻居间边的数目。所有节点聚类系数的均值表示为网络的聚类系数。

③拓扑系数(topological coefficient, T)是反映某节点的邻居间被其他节点连接在一起的比例。节点 v 的拓扑系数定义为:

$$T_v = \frac{1}{|M_v|} \sum_{t \in M_v} C_{v,t} / \min\{k_v, k_t\} \quad (3)$$

方程式(3)中 $C_{v,t}$ 表示与节点 v 和节点 t 都相连的节点数, $|M_v|$ 为所有与节点 v 分享邻居节点的节点集合大小(模), k_v, k_t 分别为节点 v 和节点 t 的连通度。

④最短路径长度(shortest paths length)定义为连通任意两个节点的最短边数,网络中所有节点对的路径长度平均值称为网络特征路径长度。

⑤节点介数(Betweenness)反映的是相应节点在整个网络中的作用和影响力,定义为网络中所有的最短路径经过本节点的数量比例。

1.7 挖掘 BPD 核心基因

假定随机网络中基因节点的连通度(x)的分布符合泊松分布^[14],我们有:

$$P(x > t) = 1 - P(x \leq t) = 1 - \sum_{k=0}^t \frac{\lambda^k}{k!} e^{-\lambda} \quad (4)$$

方程式(4)中 $\lambda = 2m/n$, t 为某一指定的连通度。在蛋白质互作网络等生物分子网络中,支持生命基本活动的必需基因或其翻译产物在网络中出现的频率显著高于一般节点。因此核心致病基因定义为其 P 值小于经 Bonferroni 校正的显著性水平。

2 结 果

2.1 基于数据库知识整合的 SNPs 筛选

经过 Hardy-Weinberg 平衡检验、数据缺失

率、MAF 以及 LD 等条件筛选之后,SNPs 数目降至 212 730 个。有近一半 SNP(98 157 个)映射到基因中。与 PPI 数据库里的基因进行交集分析,获得 6 476 个基因,对应 49 917 个 SNP。

2.2 基因网络的构建与网络拓扑性质分析

通过基于单体型 logistic 回归分析,发现 6 759 对统计显著的 SNP 互作对子(Bonferroni 校正后 $P < 0.05$),映射到 3 841 个基因。以此数据建立了 BPD 的等权重无向基因网络。通过网络拓扑性质计算得到平均连通度 $\lambda=3.743$, 聚类系数 $CC=0.033$,网络特征路径长度为 5.184。进一步对该网络几个拓扑属性(拓扑系数、最短路径长度、连通度和介数)的分布进行分析,结果如图 1 所示。从图 1A 可以看出,BPD 网络的连通度分布呈现出明显的幂律分布特点,经过数据拟合得到一条曲线($y = 2524x^{-1.934}$),具有很好的拟合性(决定系数 $R^2 = 0.945$)。如图 1B 中所示,BPD 致病基因网

络中出现频率最高的最短路径为 5 ~ 6 之间,说明了该网络不符合随机网络模型,是一种具备复杂网络“小世界(small-world)特性”的无标度(scale-free)网络^[15],即具备生物分子网络的结构特点。网络节点与其他节点分享的邻居数与其拓扑系数呈现明显的反比关系(图 1C),但与节点介数呈现正相关趋势(图 1D),反映了 BPD 致病基因网络的基本拓扑特性。

在多重检验校正的显著性水平下($\alpha = 0.05 / 3\ 841 = 1.3 \times 10^{-5}$),应用基于泊松分布的统计检验,确定了 115 个核心基因,连通度均大于 16。通过 HuGE Navigator 数据库对比发现,在 29 个连通度高于 30 的核心基因中,有 12 个 (PRKCA, EGFR, ESR1, ATXN1, FYN, CREBBP, TP53, AKT1, CSNK2A1, DLG1, PTN 和 LYN)重复了以往文献的报道,并且除了与 BPD 相关之外,大部分基因也曾被报道过与其他神经、精神疾病(如精

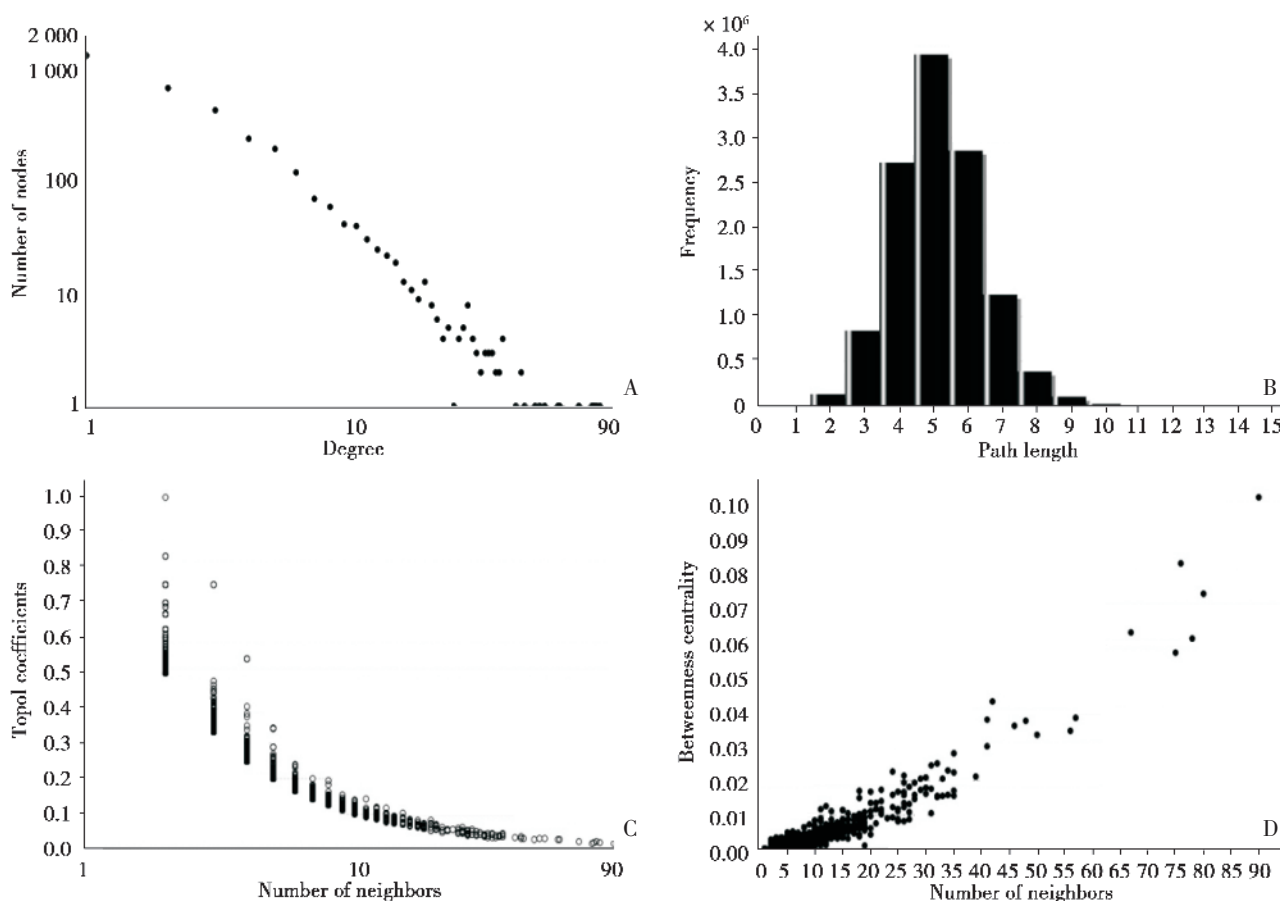


图 1 BPD 网络的拓扑属性

Fig.1 Topological properties of the gene network for BPD

A: Node degree; B: Shortest path length; C: Topological coefficients; D: Betweenness.

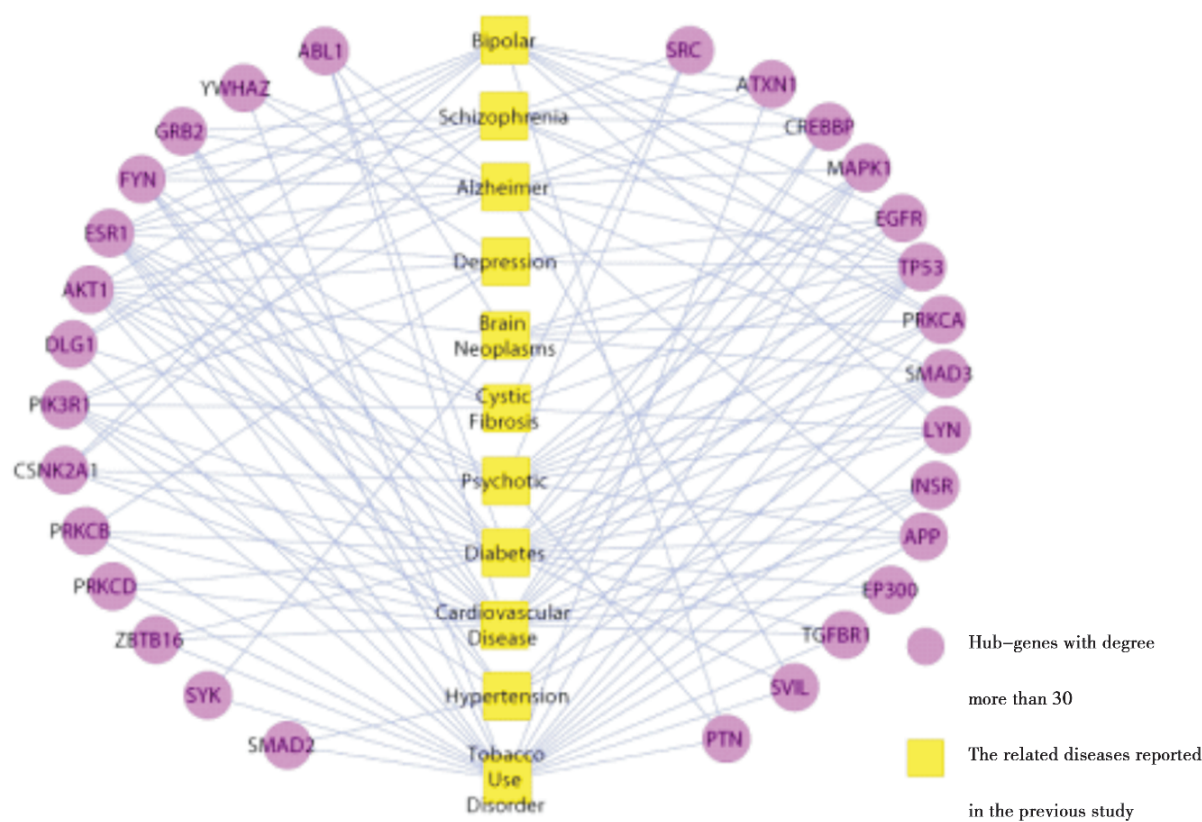


图 2 网络中连通度高于 30 的核心致病基因与其被报道过的相关疾病
Fig.2 Hub-genes with degree more than 30 and their reported associated diseases

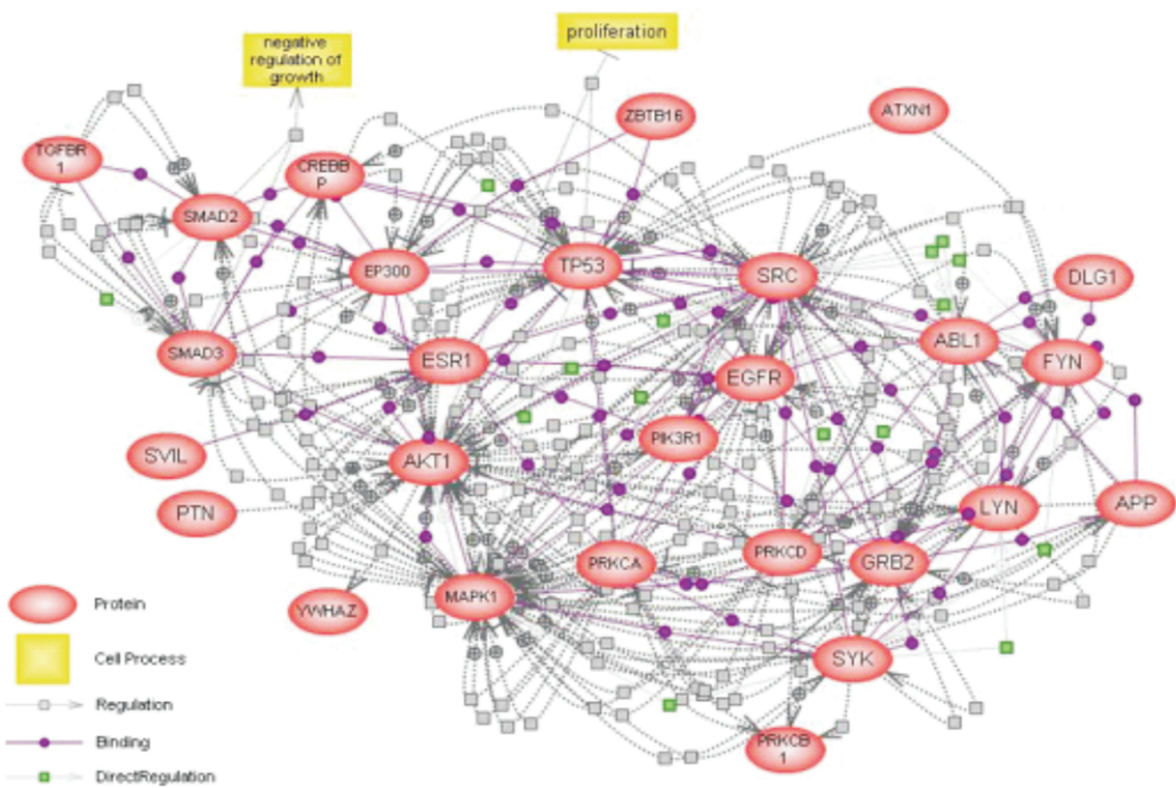


图 3 核心致病基因网络图
Fig.3 The gene network involving the hub genes

神分裂症、自闭症、抑郁症、阿尔茨海默病或囊性纤维病等)以及糖尿病、心血管疾病以及烟草滥用等相关,说明本研究提出分析方法的有效性。其余的 17 个核心基因(SMAD3, SRC, GRB2, PIK3R1, ZBTB16, ABL1, APP, EP300, TGFBR1, SYK, YWHAZ, INSR, MAPK1, PRKCB, PRKCD, SMAD2 和 SVIL)虽然没有被报道过,但是大多数曾被报道过与其他脑部疾病(如脑肿瘤、偏头疼)、精神或神经疾病(如精神分裂症、自闭症、阿尔茨海默病或囊性纤维病等)以及心血管疾病、糖尿病等疾病相关联。图 2 中展示了连通度大于 30 的核心风险基因与 BPD 及相关疾病的关系,29 个圆圈代表了连通度大于 30 的核心基因,中间 11 个正方形分别代表 11 种被报道过的与核心基因相关的复杂疾病,连线表示基因与相关疾病之间的对应关系。

为了进一步了解这些核心基因的分子调控机制,运用 Pathway Assist 软件和数据库绘制出核心致病基因网络图(图 3)。可见,TGFBR1 调控了 SMAD2 与 SMAD3,三者共同控制生长负调控过程;AKT1 与 TP53 结合调控细胞增殖过程;网络中 AKT1、EGFR、SRC 和 MAPK1 均受到周围多个基因的直接调控,说明这几个基因在疾病发生的过程中可能发挥了核心的作用。

3 讨 论

通过分析核心基因的分子生物学功能,可以知道:APP 基因与 5-羟色胺能突触传递及阿尔茨海默病的发生发展密切相关^[16],GRB2、PIK3R1、ABL1 和 PRKCD 在神经营养因子信号转导通路中扮演着重要的角色^[17]。ZBTB16 是影响中枢神经系统的发育重要效应基因之一,中枢神经递质代谢异常、相应受体功能改变、大脑神经突触间隙 5-羟色胺(5-HT)等神经递质水平异常都能诱导 BPD 的发生^[18]。 γ -氨基丁酸(GABA)是中枢神经系统抑制性神经递质, SRC 与 PRKCB 基因的突变会调控 GABA 能突触通路下调血浆和脑脊液中的 GABA 含量,从而诱发 BPD^[19]。Mayor 等^[20]发现 TGFBR1、MAPK1 和 PRKCB 等基因共同参与的 MAPK 信号通路激活后,可导致脑颗粒细胞发生凋亡。这些证据都可以很好地解释核心基因的致病机制,也验证了我们的预测结果是较为准确的,

可以为以后的研究提供新的方向。

本文提出了一种基于数据库知识整合的数据降维方法进行 BPD 全基因组关联研究的新思路。根据复杂疾病的生物学机制以及结合蛋白互作的本质,从致病基因网络整体出发并通过网络的拓扑性质寻找核心致病风险基因,不仅在计算过程中避开了单位点分析所带来的弊端,弥补了现有方法的不足,而且使得统计分析的结果更加合理准确。既能够重复以往前人的调查研究结果,又提出了一些可能影响 BPD 致病易感性基因,这将有助于为后续的实验室研究提供指导与帮助。

网络拓扑性质的计算分析,量化了各基因在网络中的重要性以及相互连接的紧密性,从而也验证了构建所得的基因网络并非随机模型,而是符合生物分子网络结构的无标度网络。进一步依据分子网络的生物学性质以及拓扑性质分析核心风险基因,为揭开 BPD 遗传机制提供了更为明确的方向。实验结果发现,网络中的中心基因有很多是已报道过的 BPD 风险基因,另外一些中心基因虽然未被报道过,但是大部分与脑部疾病、精神疾病或神经系统疾病相关,所以本研究具有前瞻性,可以作为一种有效的预测工具。但需要指出的是,由于本研究采用数据库知识引导进行数据降维的策略,一定程度上依赖于数据库的信息量,所以难免会错失一些重要的信息。目前只能尽可能地搜索信息更全的数据库,设法将损失的信息降到最低。另一方面,数据本身的缺陷也会影响研究结果,如疾病诊断的同质性、没有疾病亚型分层及环境变量数据的缺失等,在以后的研究中我们应采用信息更全面的数据进行可重复性验证。我们还将考虑将细胞信号通路的信息融合到网络中,从而能够更加详细地研究 BPD 的分子代谢和调控机制。

(致谢:衷心感谢中山大学公共卫生学院左晓宇同学为本研究提出的宝贵意见;感谢中山大学数学与计算科学学院的赵忠同学在数据分析等方面为我们提供的支持和帮助)

参考文献:

- [1] Kessler RC, Berglund P, Demler O, et al. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication [J]. Arch Gen Psychiatry, 2005, 62(6): 593-602.

- [2] Kessler RC, Akiskal HS, Ames M, et al. Prevalence and effects of mood disorders on work performance in a nationally representative sample of U.S. workers[J]. *Am J Psychiatry*, 2006, 163(9): 1561–1568.
- [3] Smoller JW, Finn CT. Family, twin, and adoption studies of bipolar disorder[J]. *Am J Med Genet C Semin Med Genet*, 2003, 123C(1): 48–58.
- [4] Purcell, Shaun M, Wray NR, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder[J]. *Nature*, 2009, 460(7256): 748–752.
- [5] Risch N, Merikangas K. The future of genetic studies of complex human diseases [J]. *Science*, 1996, 273(5281): 1516–1517.
- [6] Risch NJ. Searching for genetic determinants in the new millennium[J]. *Nature*, 2000, 405(6788): 847–856.
- [7] Tavazoie S, Hughes JD, Campbell MJ, et al. Systematic determination of genetic network architecture [J]. *Nat Genet*, 1999, 22(3): 281–285.
- [8] Akutsu T, Miyano S, Kuhara S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model[J]. *Pac Symp Biocomput*, 1999: 17–28.
- [9] Friedman N, Linial M, Nachman I, et al. Using Bayesian networks to analyze expression data [J]. *J Comput Biol*, 2000, 7(3–4): 601–620.
- [10] WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls[J]. *Nature*, 2007, 447(7145): 661–678.
- [11] Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data [J]. *Am J Hum Genet*, 2003, 73(6): 1316–1329.
- [12] Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes [J]. *Am J Hum Genet*, 1995, 56(3): 799–810.
- [13] Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization [J]. *Nat Rev Genet*, 2004, 5(2): 101–113.
- [14] Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network [J]. *Bioinformatics*, 2006, 22(22): 2800–2805.
- [15] Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks[J]. *Nature*, 1998, 393(6684): 440–442.
- [16] Bockaert J, Claeysen S, Becamel C, et al. Neuronal 5-HT metabotropic receptors: fine-tuning of their structure, signaling, and roles in synaptic modulation [J]. *Cell Tissue Res*, 2006, 326(2): 553–572.
- [17] Arevalo JC, Wu SH. Neurotrophin signaling: many exciting surprises! [J]. *Cell Mol Life Sci*, 2006, 63(13): 1523–1537.
- [18] Suzuki K, Kusumi I, Sasaki Y, et al. Serotonin-induced platelet intracellular calcium mobilization in various psychiatric disorders: is it specific to bipolar disorder? [J]. *J Affect Disord*, 2001, 64(2–3): 291–296.
- [19] Benes FM, Berretta S. GABAergic interneurons: implications for understanding schizophrenia and bipolar disorder[J]. *Neuropsychopharmacology*, 2001, 25(1): 1–27.
- [20] Mayor F, Jurado-Pueyo M, Campos PM, et al. Interfering with MAP kinase docking interactions: implications and perspective for the p38 route [J]. *Cell Cycle*, 2007, 6(5): 528–533.

(编辑 刘清海)