Pathway-based analysis for genome-wide association studies of schizophrenia to provide new insight in schizophrenia study

ZHANG KunLin[1], ZHANG LiuYan[1,2], ZHANG WeiNa[1,2] & WANG Jing[1*]

[1] Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China;
[2] Graduate University of Chinese Academy of Sciences, Beijing 100049, China

Schizophrenia (SZ) is an inheritable complex mental disease. There have been several genome-wide association studies (GWASs) of SZ to identify novel genetic susceptibility factors. To further interpret SZ GWASs, pathway-based analysis (PBA), which considers the combined effect of variants and identifies pathways associated with traits, provides a feasible solution to discover the biological function and mechanism of SZ. Furthermore, to investigate the common pathways between SZ and bipolar disorder (BD) will help explore common mechanism between psychiatric phenotypes. We performed a PBA, called improved gene set enrichment analysis (i-GSEA), on 3 independent GWASs of SZ to identify pathways associated with SZ. The results were further compared to the BD-associated pathways identified by i-GSEA for 2 BD GWASs and from literature reports. Our analysis identified a highly statistically significant association between SZ and pathway 'substrate specific channel activity' in all 3 SZ GWASs (false discovery rate (FDR) < 0.05). This association has not been reported elsewhere before. This pathway was also identified by PBA for 2 independent BD GWASs. Our results suggest that pathway 'substrate specific channel activity' is statistically significantly associated with SZ, and SZ and BD share the common biological function and mechanism represented by this pathway.

schizophrenia (SZ), genome-wide association study (GWAS), pathway-based analysis (PBA), bipolar disorder (BD)

Schizophrenia (SZ) is a complex mental disease, characterized by hallucinations, delusions and cognitive deficits and caused by both genetic and environmental factors and their interactions. SZ has a global lifetime prevalence about 0.4%–1% [1,2], with heritability estimated at approximately 80%–85% [3]. However, the susceptibility genes or biological mechanisms for SZ are still not well-known.

Genome-wide association study (GWAS), which utilizes genome-wide single nucleotide polymorphism (SNP) array to identify genetic susceptibility through detecting the association between genotype frequency and trait [4], provides an opportunity to detect the effects of common variants with greater power than linkage analysis. According to the collection of GWAS Catalog (www.genome.gov/gwastudies) [5], there have been 10 SZ GWASs till 2010, which were mainly based on European ancestry population. However, the first 6 SZ GWASs did not find any genome-wide significant or meaningful variant which is expected to have the potential association with SZ [6–11]. The reasons are probably these studies have been performed using pooled DNA, ethnically heterogeneous samples or small samples sizes (there is no study including more than 1000 cases), which reduced the power of GWAS to detect the effects of common SNPs. After increasing the sample size and improving the experimental design and analytical strategy, the SGENE consortium, the International Schizophrenia Consortium (ISC) and the Molecular Genetics of Schizophrenia consortium (MGS) found several genome-wide significant SNPs in the major histocompatibility complex (MHC) region on chromosome 6 [12–14]. Besides this, several significant variants also indicate the association between SZ and the genes involved in brain development, memory and cogni-

*Corresponding author (email: wangjing@psych.ac.cn)

tion. However, these GWAS analyses focused on single SNP/gene and identified only a number of the most significant SNPs, which accounted for a small proportion of the genetic variants, so they offered limited understanding of schizophrenia. Thus, the method for detecting the combined genetic effect of SZ is expected to be applied to SZ GWASs for further GWAS data interpretation and mechanism study of SZ.

Pathway-based analysis (PBA) which was originally developed for analyzing gene expression data [15] has been introduced to GWAS data analysis to detect the association between pathway and trait [16–18]. It has been proved to be a feasible solution to interpret GWAS result and to promote the analytical level from SNP/gene to pathway. An important PBA approach is the gene set enrichment analysis (GSEA) which technically evaluates whether the genes sharing a biological function is enriched at the top of a ranked genome-wide gene list [16]. In our previous work, we designed and implemented a novel PBA method, called *i*-GSEA (improved gene set enrichment analysis). *i*-GSEA employs the analytical framework of GSEA while it emphasizes on identifying the pathways containing high proportion of significant genes. Our study has shown that *i*-GSEA is a powerful PBA approach to detect the disease-related pathways reflecting the combined genetic effect of complex disease with high sensitivity [19]. To perform PBA (i.e. *i*-GSEA) on SZ GWASs will help discover pathways associated with SZ to provide new insight in SZ study. In addition, since SZ is assumed to share some common genetic factors with bipolar disorder (BD) [20], applying PBA on BD GWASs to investigate the pathways associated with both SZ and BD will help explore common mechanism between psychiatric phenotypes (SZ and BD).

In this study, we performed the PBA (*i*-GSEA) on 3 independent SZ GWASs, and then we investigated these pathways by comparing them with the BD-associated pathways identified by PBA (*i*-GSEA) for 2 BD GWASs and from 7 PBA reports from literatures. We identified that pathway 'substrate specific channel activity' is associated with SZ with highly statistical significance (false discovery rate (FDR) < 0.05 in all 3 SZ GWASs). This pathway was also detected by PBA for 2 independent BD GWASs.

1   Materials and methods

1.1   Materials

We requested and obtained 2 SZ GWAS SNP *P*-value datasets, namely Genetic Association Information Network (GAIN) European ancestry (GAIN_EA, Accession No. pha002857.v1.p1) and GAIN African American ancestry (GAIN_AA, Accession No. pha002859.v1.p1), from dbGaP (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap). Data access was approved by the GAIN DAC. The third SZ GWAS SNP *P*-value data set from Duke Study European ancestry (DS_EA) samples were obtained directly from ref. [11]. These 3 datasets are the only data that allow public access and none of them have identified SNP(s) with genome-wide significance. In the same way, we requested and obtained 2 BD GWAS data, namely GAIN European ancestry (GAIN_EA, Accession No. pha002858.v1.p1) and GAIN African American ancestry (GAIN_AA, Accession No. pha002863.v1.p1), from dbGaP. The details of the 3 GWAS datasets of SZ and 2 GWAS datasets of BD are shown in Table 1.

Pathway dataset, which serves as the search space of our PBA, consists of pathways from BioCarta (http://www. biocarta.com/), KEGG [21], and gene ontology (GO) [22] biological process domain and molecular function domain (we used curated GO terms from MSigDB [23] plus level 4 GO terms). In order to avoid over large or over small search space and functional category, only the pathways with size (number of genes involved) more than 20 and less than 200 were used in our PBA analysis.

Other PBA results of BD, which were used for a comprehensive comparison between our PBA results of SZ and BD-associated pathways, were collected from refs. [19, 24–32], as detailed in Table 2.

1.2   Methods

To analyze the 3 GWAS datasets for SZ and the 2 GWAS datasets for BD, for each dataset, we firstly retrieved SNP and gene localizations on human chromosomes from the Ensembl database [33]. Then we associated each SNP with its nearest gene with distance less than 500 kb [16]. Second,

Table 1   The detailed information of different GWASs of SZ and BD[a]

| GWAS dataset (disease name)	Ethnicity	Number of cases/controls	Number of genotyped SNPs
GAIN_EA (SZ)	European ancestry	1351/1378	over 729454
GAIN_AA (SZ)	African American ancestry	1195/954	over 845814
DS_EA (SZ)	European ancestry	900/877	over 312565
GAIN_EA (BD)	European ancestry	1001/1034	over 728222
GAIN_AA (BD)	African American ancestry	362/671	over 845814

a) GAIN_EA, Genetic Association Information Network European ancestry; GAIN_AA, Genetic Association Information Network African American ancestry; DS_EA (SZ), Duke Study European ancestry for SZ (ref. [11]).

Table 2   Summary of the available PBA results for BD GWASs[a)]

GWAS dataset	PBA approach	Major finding	Reference
WTCCC	*i*-GSEA	glutathione metabolism	[19]
WTCCC	hypergeometric test	heparan sulfate, Niacin-HDL metabolism, adrenergic receptor-mediated cytoskeletal remodeling pathways	[24]
WTCCC	Prioritizing Risk Pathways fusing SNPs and pathways (PRP)	purine metabolism, phosphatidylinositol signaling system, MAPK signaling pathway, p53 signaling pathway, arachidonic acid metabolism, WNT signaling pathway, *etc*	[25]
WTCCC	hypergeometric test (Fisher's exact test), Sidak's method, Simes' method, Simes/FDR method	Inositol metabolism, Glycine, serine and threonine metabolism, N-Glycan biosynthesis, Chondroitin sulfate biosynthesis, Fatty acid metabolism, Nucleotide sugars metabolism, Chondroitin sulfate biosynthesis, Heparan sulfate biosynthesis, *etc*	[26]
WTCCC	SNP ratio test	cell adhesion molecule (CAM)	[27]
WTCCC + NIMH	Fisher exact test	ion channel activity, synaptic neurotransmission	[28]
WTCCC + STEP-UCL + ED-DUB-STEP2	ALIGATOR (Association LIst Go AnnoTatOR)	transcription and cellular activity, including that which occurs via hormonal action	[29]

a) WTCCC, Wellcome Trust Case Control Consortium (ref. [30]); NIMH, National Institutes of Mental Health (ref. [31]); STEP-UCL, Systematic Treatment Enhancement Program for Bipolar Disorder-University College London; ED-DUB-STEP2, University of Edinburgh-Trinity College Dublin-Systematic Treatment Enhancement Program for Bipolar Disorder additional (ref. [32]).

the PBA, i.e. *i*-GSEA as reported in our recent study [19], was applied to the SNP *P*-values of the 3 SZ GWASs respectively, by using the pathway datasets described above as search space. The key running procedures of *i*-GSEA are as follows: (1) The lowest *P*-value of the SNPs of a gene is utilized to represent the gene. (2) For each pathway, enrichment score (ES, a Kolmogorov-Smirnov like statistics with weight 1), which reflects the trend that genes of the pathway held better associations than genome-wide genes, is calculated based on ranked gene list. (3) A significant proportion ratio, which emphasizes the relative proportion of significant genes (defined as genes mapped with at least 1 SNP with *P*-value < 0.05), is multiplied to ES to obtain significance proportion based ES (SPES). (4) SNP label permutation and normalization are performed to generate the null distribution of SPES and correct gene variation and gene set variation. Let u be permutation and S be pathway. For a specific permutation u^* and a specific pathway S^*, $\text{SPES}(u^*, S^*)$ is computed by repeating (1)–(3). Normalized $\text{SPES}(S^*)$ (i.e. $\text{NSPES}(S^*)$) equals to $(\text{SPES}(S^*) - m)/sd$, where m is the mean of $\text{SPES}(u, S^*)$ and sd is the standard deviation of $\text{SPES}(u, S^*)$. $\text{NSPES}(u^*, S^*)$ is calculated by using the same way as $\text{NSPES}(S^*)$. (5) Nominal *P*-value is calculated and false discovery rate (FDR) is used for multiple testing correction [34]. For a specific pathway S^*, nominal *P*-value = (number of $\text{NSPES}(u, S^*)$ which are greater than or equal to $\text{NSPES}(S^*)$)/(number of $\text{NSPES}(u, S^*)$). FDR = (percentage of $\text{NSPES}(u, S)$ which are greater than or equal to $\text{NSPES}(S^*)$)/(percentage of $\text{NSPES}(S)$ which are greater or equal to $\text{NSPES}(S^*)$). We kept those pathways with FDR < 0.05 as candidate pathways associated with SZ. Figure 1 shows the key procedures of *i*-GSEA. Finally, we compared the *i*-GSEA results of SZ with BD-associated pathways (our *i*-GSEA results of BD and other PBA results from literatures) to study the common function and mechanism between SZ and BD.
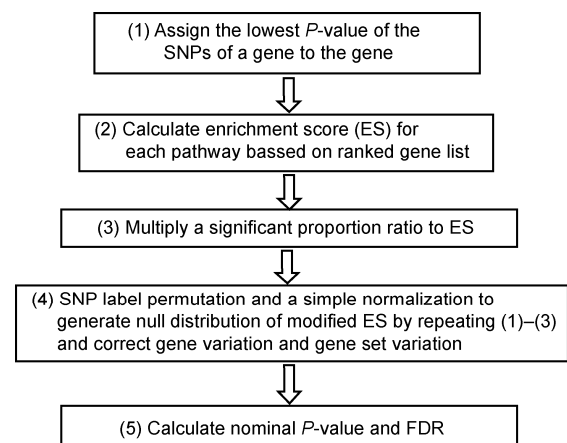
Figure 1   Key procedures of *i*-GSEA.

2 Results

2.1 Overview of the results

In PBA for SZ GAIN_EA data, 674868 SNPs were mapped to 17198 genes, 980 pathways were selected, and 36 pathways with FDR < 0.05; in PBA for SZ GAIN_AA data, 782587 SNPs were mapped to 17440 genes, 991 pathways were selected, and 5 pathways with FDR < 0.05; in PBA for SZ DS_EA data, 288615 SNPs were mapped to 16697 genes, 972 pathways were selected, and 19 pathways with FDR < 0.05. Our analysis identified the highly statistically significant (FDR < 0.05) association between schizophrenia and pathway 'substrate specific channel activity' in all 3 GWASs (FDR = 0.034, 0.047, and 0.016 in GAIN_EA, GAIN_AA, and DS_EA respectively). Immunity-related pathways with highly statistical significance were also identified in all the 3 GWASs of SZ by our PBA analysis. Since they have been well reported before, we will not elaborate it in this paper.

In PBA for BD GAIN_EA data, 673770 SNPs were

mapped to 17186 genes, 981 pathways were selected and 60 pathways with FDR < 0.05; in PBA for BD GAIN_AA data, 782587 SNPs were mapped to 17440 genes, 991 pathways were selected and 3 pathways with FDR < 0.05. There is no pathway with FDR < 0.05 in both GWASs.

Comparison between SZ-associated pathway 'substrate specific channel activity' by PBA and BD-associated pathways show that pathway 'substrate specific channel activity' is shared by both SZ and BD, which indicate the possible common function and mechanism between the 2 diseases.

2.2 Pathway 'substrate specific channel activity'

The pathway 'substrate specific channel activity' (GO Accession: GO: 0022838) is defined as energy-independent facilitated diffusion catalysis that is mediated by passage of a specific solute through a channel or transmembrane aqueous pore. This pathway consists of 3 sub-pathways (or called child nodes of GO tree), namely 'ion channel activity' (GO: 0005216), 'urea channel activity' (GO: 0015265), and 'water channel activity' (GO: 0015250). The pathway 'substrate specific channel activity' describes a specific kind of channel activity, which covers ion channel activity. It is well known that ion channels play extremely important roles in potential based neuron function, so the abnormality of this activity may be a possible risk factor of SZ.

The pathway 'substrate specific channel activity' is also presented in our result of PBA for BD GWAS (FDR = 0.042, for BD GWAS of GAIN_EA) and its sub-pathway 'ion channel activity' is presented in the result of PBA for BD GWAS performed by Askland et al. [28], which analyzed the Wellcome Trust Case Control Consortium (WTCCC) data [30] and National Institutes of Mental Health (NIMH) data [31] (Bonferroni-adjusted *P*-value = 0.0064 in WTCCC and Bonferroni-adjusted *P*-value = 0.0035 in NIMH, respectively). Thus, the pathway 'substrate specific channel activity' was identified by 2 PBA analyses corresponding to 2 independent GWASs of BD.

The significant PBA result of SZ and its shared common pathway with BD are shown in Table 3.

3 Discussion

As the GWASs for SZ identified only a few number of SNPs/genes and offer limited understanding of the disease,

PBA is important to further interpret SZ GWASs to detect functional units associated with SZ. In our PBA for different SZ GWASs, the pathway 'substrate specific channel activity' was identified. Pathway 'substrate specific channel activity' was not presented in single SNP/gene based analysis and its association with SZ was first reported by us. This shows that our analysis has the power of interpreting the GWASs of SZ to further identify disease-related mechanism and provide new insight into the study of SZ. It should be noted that the PBA approach, *i*-GSEA, does not take into account the linkage disequilibrium (LD) of SNPs, since LD can only be analyzed when genotype data is available (for most published GWAS investigations, genotype data is usually not available). It assumes that all the input SNPs are independent (i.e. there is no LD between them) while performing permutation. This may lead to some bias as SNPs tend to be non-independent. So for GWAS with genotype data available, to input SNPs not in LD (say $r^2 < 0.2$) is recommended to reduce the possibility of biased results.

The pathway 'substrate specific channel activity' found in this report is statistically replicable and across ethnic groups because it's presented in 3 PBA results for 3 independent SZ GWASs (including 2 GWASs for European ancestry population and 1 GWAS for African American ancestry population). Recently, a PBA analysis, which combined GSEA and Fisher exact test, for 1 single SZ GWAS (GAIN_EA) has found that the metabolism of glutamate, the process of apoptosis, inflammation, and immune system could be associated with SZ [35]. The other PBA analysis of SZ, which utilized SNP ratio test, for 2 SZ GWASs, namely GAIN_EA and ISC (International Schizophrenia Consortium), identified that pathway 'cell adhesion molecule (CAM)' (an immune-related pathway) is associated with SZ [27]. This research has also found that CAM is also associated with BD (by analyzing WTCCC BD GWAS [30]). There are 2 reasons may lead to the different results between these 2 researches and our research. One is that the above 2 researches only focused on the population with European ancestry while our work focused on populations both with European ancestry and African ancestry. The other reason is that different researches utilized different methods of PBA with different parameters. Our result has been proved to be solid because it is statistically significant in all 3 independent SZ GWASs corresponding to 2 different ethnic groups. Apparently for the same trait, different ethnic groups should share the same

Table 3   The significant result of pathway 'substrate-specific channel activity' in PBA analyses for SZ and BD GWASs

GWAS dataset (disease name)	PBA approach	*P*-value	FDR
GAIN_EA (SZ)	*i*-GSEA	< 0.001	0.034
GAIN_AA (SZ)	*i*-GSEA	< 0.001	0.047
DS_EA (SZ)	*i*-GSEA	< 0.001	0.016
GAIN_EA (BD)	*i*-GSEA	< 0.001	0.042
WTCCC + NIMH (BD)	Fisher exact test	1.27E-5[a] (in WTCCC) 6.99E-6[a] (in NIMH)	0.0064[b] (in WTCCC) 0.0035[b] (in NIMH)

a) For 'ion channel activity', the sub-pathway of 'substrate-specific channel activity'; b) Bonferroni-adjusted *P*-value for sub-pathway 'ion channel activity'.

biological functions and mechanisms.

It is the first time that the association between pathway 'substrate specific channel activity' and SZ is identified, and the association is statistically significant and outstandingly solid. It has been known that SZ and BD share some common genetic factors [12]. Our study further indicates that they share some common biological mechanism, 'substrate specific channel activity', and it is not found by the single SNP/gene based analysis of GWAS. This novel result provides the new evidence for the shared biological mechanisms between SZ and BD.

4 Conclusions

Our results suggest that the pathway 'substrate specific channel activity' is associated with SZ, and SZ and BD share this common biological function and mechanism.

1　Perala J, Suvisaari J, Saarni S I, et al. Lifetime prevalence of psychotic and bipolar I disorders in a general population. Arch Gen Psychiatry, 2007, 64: 19–28

2　Saha S, Chant D, Welham J, et al. A systematic review of the prevalence of schizophrenia. PLoS Med, 2005, 2: e141

3　Cardno A G, Gottesman I I. Twin studies of schizophrenia: From bow-and-arrow concordances to star wars Mx and functional genomics. Am J Med Genet, 2000, 97: 12–17

4　McCarthy M I, Abecasis G R, Cardon L R, et al. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nat Rev Genet, 2008, 9: 356–369

5　Hindorff L A, Sethupathy P, Junkins H A, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA, 2009, 106: 9362–9367

6　Lencz T, Morgan T V, Athanasiou M, et al. Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. Mol Psychiatry, 2007, 12: 572–580

7　O'Donovan M C, Craddock N, Norton N, et al. Identification of loci associated with schizophrenia by genome-wide association and follow-up. Nat Genet, 2008, 40: 1053–1055

8　Shifman S, Johannesson M, Bronstein M, et al. Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women. PLoS Genet, 2008, 4: e28

9　Sullivan P F, Lin D, Tzeng J Y, et al. Genomewide association for schizophrenia in the CATIE study: Results of stage 1. Mol Psychiatry, 2008, 13: 570–584

10　Kirov G, Zaharieva I, Georgieva L, et al. A genome-wide association study in 574 schizophrenia trios using DNA pooling. Mol Psychiatry, 2009, 14: 796–803

11　Need A C, Ge D, Weale M E, et al. A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet, 2009, 5: e1000373

12　Purcell S M, Wray N R, Stone J L, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature, 2009, 460: 748–752

13　Shi J, Levinson D F, Duan J, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature, 2009, 460: 753–757

14　Stefansson H, Ophoff R A, Steinberg S, et al. Common variants conferring risk of schizophrenia. Nature, 2009, 460: 744–747

15　Nam D, Kim S Y. Gene-set approach for expression pattern analysis. Brief Bioinform, 2008, 9: 189–197

16　Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet, 2007, 81: 1278–1283

17　Cantor R M, Lange K, Sinsheimer J S. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet, 2010, 86: 6–22

18　Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet, 2010, 11: 843–854

19　Zhang K, Cui S, Chang S, et al. i-GSEA4GWAS: A web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. Nucleic Acids Res, 2010, 38(Suppl): W90–W95

20　Craddock N, O'Donovan M C, Owen M J. Genes for schizophrenia and bipolar disorder? Implications for psychiatric nosology. Schizophr Bull, 2006, 32: 9–16

21　Kanehisa M, Goto S, Furumichi M, et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res, 2010, 38: D355–D360

22　Ashburner M, Ball C A, Blake J A, et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, 25: 25–29

23　Subramanian A, Tamayo P, Mootha V K, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA, 2005, 102: 15545–15550

24　Torkamani A, Topol E J, Schork N J. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics, 2008, 92: 265–272

25　Chen L, Zhang L, Zhao Y, et al. Prioritizing risk pathways: A novel association approach to searching for disease pathways fusing SNPs and pathways. Bioinformatics, 2009, 25: 237–242

26　Peng G, Luo L, Siu H, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet, 2010, 18: 111–117

27　O'Dushlaine C, Kenny E, Heron E, et al. Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. Mol Psychiatry, 2011, 16: 286–292

28　Askland K, Read C, Moore J. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. Hum Genet, 2009, 125: 63–79

29　Holmans P, Green E K, Pahwa J S, et al. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet, 2009, 85: 13–24

30　The Wellcome Trust Case Control Consortium. Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. Nature, 2007, 447: 661–678

31　Sklar P, Smoller J W, Fan J, et al. Whole-genome association study of bipolar disorder. Mol Psychiatry, 2008, 13: 558–569

32　Ferreira M A, O'Donovan M C, Meng Y A, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet, 2008, 40: 1056–1058

33　Flicek P, Aken B L, Ballester B, et al. Ensembl's 10th year. Nucleic Acids Res, 2010, 38: D557–D562

34　Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics, 2003, 19: 368–375

35　Jia P, Wang L, Meltzer H Y, et al. Common variants conferring risk of schizophrenia: A pathway analysis of GWAS data. Schizophr Res, 2010, 122: 38–42