Literature Data Mining and Enrichment Analysis on Top 235 Genes for Attention Deficit Hyperactivity Disorder

Shunan Li[1], Karan Kapoor[2], Lydia C Manor[3*]

[1] Vanderhouwen & Associates, 6342 SW Macadam Ave, Portland, OR 97239, USA; [2] Center for Molecular Biophysics, University of Tennessee / Oak Ridge National Laboratory, 1 Bethel Valley Road, Bldg 2040, Oak Ridge , TN 37830, USA; [3] American Informatics Consultant LLC, Rockville, MD, 20852, USA.

***Correspondence:** Dr. Lydia C Manor, American Informatics Consultant LLC, Rockville, MD, 20852, USA. Email: l.manor@gousinfo.com.

ABSTRACT

Background: Attention deficit hyperactivity disorder (ADHD) is a psychiatric disorder of the neuro-developmental type, marked by an ongoing pattern of inattention or hyperactivity/impulsivity, which interferes with functioning or development. The disorder affects approximately 5-7 % children and 2-5 % of adults worldwide. Numerous studies have indicated that genetic factors predominate the causes for ADHD. Nevertheless, no systematic study has summarized these findings and provided an objective and complete list of genes with a reported association to ADHD.

Methods: Literature and enrichment metrics analyses were used to discover genes of specific significance associated with ADHD. We conducted a literature data mining (LDM) of over 2,410 articles covering publications from Jan. 1988 to Apr. 2016, where 235 genes were reported to be associated with the disease. Then we performed a gene set enrichment analysis (GSEA) and a sub-network enrichment analysis (SNEA) to study the functional profile and pathogenic significance of these genes associated with ADHD. Lastly, we performed a network connectivity analysis (NCA) to study the associations between the reported genes.

Results: 181/235 genes enriched 100 pathways ($p<1.1e-007$), demonstrating multiple associations with ADHD. Twelve genes were discovered to be associated with ADHD, in terms of both functional diversity and replication frequency, including SLC6A3, DRD4, BDNF, DRD2, HTR2A, DBH, HTR1B, DRD5, GRM7, DRD3, TH and GRIN2A. In addition, one novel gene, SHANK2, was suggested worthy of further study. Moreover, SNEA and NCA results indicated that many of these genes form a functional network, playing roles in the pathogenesis of other ADHD related disorders.

Conclusion: Our results suggest that the genetic causes of ADHD are linked to a genetic and functional network composed of a large group of genes. The gene lists, together with the literature and enrichment metrics provided in this study, could serve as groundwork for further biological/genetic studies in the field.

Key Words: Attention Deficit Hyperactivity Disorder (ADHD); Enrichment Analysis; Gene

OPEN ACCESS

DOI: 10.20900/jpbs.20160008

Received: March 15, 2016

Accepted: May 18, 2016

Published: June 25, 2016

website: http://jpbs.qingres.com

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

INTRODUCTION

Attention deficit hyperactivity disorder (ADHD) is a brain disorder characterized by problems in paying attention, excessive activity, or difficulty in controlling behavior that is inappropriate for a person's age.(1) The World Health Organization estimated that it affected around 39 million people as of 2013.(2) ADHD is diagnosed approximately three times more in boys than in girls.(3) Despite being the most commonly studied and diagnosed psychiatric disorder in children and adolescents, the cause for the disorder remains unknown in majority of the cases. Numerous studies have indicated that the onset of the disorder involves an interaction between the genetic and environmental factors.(4)

There have been an increased number of articles reporting hundreds of genes/proteins related to ADHD, many of which have been suggested as potential biomarkers for the disease, such as SLC6A3 and ADRA2A. (5,6) Some of these genes (e.g., SLC6A2) have been studied in clinical trials as well.(7) Articles have also reported on the quantitative changes in gene expression in the case of ADHD.(8,9) Both increased and decreased gene expression levels/activities have been observed.(10,11) To note, many genes were also reported to influence the pathogenic development of ADHD with an unknown mechanism. (12,13)

Some recent studies have suggested a functional mechanism of a mutation that can cause ADHD. Hong et al. showed that differential expressions of Homer 1a and Homer 2a/b, a family of scaffolding proteins localized to the postsynaptic density of glutamatergic excitatory synapses, were observed in the prefrontal cortex and extended to the hippocampus. These genes have direct connections to attention and cognition, the two functions that were disturbed in ADHD. (14)

Nevertheless, no systematic analysis has been done to evaluate the quality and strength of these reported genes as a functional network/group in order to study the underlying biological processes of ADHD. In this study, instead of focusing on a specific gene, we attempt to provide a full view of the genetic-map, and use gene set enrichment analysis (GSEA), as well as a sub-network enrichment analysis (SNEA) to study the underlying functional profile of the genes identified.(15) We hypothesized that the majority of these previously reported genes, if not all of them, play roles in the development of ADHD, and that the major pathways/gene sets enriched by these genes are the ones associated with the disease.

METHODS AND MATERIALS

The workflow of the study is as follows: 1) Literature data mining (LDM) to discover gene-ADHD relationship; 2) Enrichment analysis on the identified genes to study their pathogenic significance in ADHD; 3) Literature and enrichment metrics analysis; and 4) Network connectivity analysis (NCA) to test the functional association between these reported genes.

1. Literature Data Mining and Article Selection Criterion

In this study, we performed a LDM for all articles available in the Pathway Studio database (www.pathwaystudio.com), that covered over 40 million scientific articles up until Apr. 2016, seeking the ones that reported gene-ADHD relations. The LDM was conducted by employing the finely-tuned Natural Language Processing (NLP) system of the Pathway Studio software, which has the capability of identifying and extracting relationship data from scientific literature. Only the publications containing a biological gene-ADHD interaction defined by ResNet Exchange (RNEF) data format was included (http://www.gousinfo.com/). Results were presented with a full list of genes names, the information of the underlying articles, and the metrics scores, which are described below.

2. Literature Metrics Analysis

For literature metrics analysis, we proposed two scores for each gene-disease relationship.We define the reference number underlying a gene-disease relationship as the gene's reference score (RScore), given by Eq. (1).

$$RScore = n \quad (1)$$

where, n is the total number of references supporting a gene-disease relation. We also define the earliest publication age of a gene-disease relationship as the gene's age score (AScore), given by Eq. (2).

$$AScore=\max_{(1\le i\le n)} ArticlePubAge_i \quad (2)$$

where, n is the total number of references supporting a gene-disease relationship, and a Article publication age (ArticlePubAge), given by Eq. (3).

ArticlePubAge=Current date - Publication date + 1 (3)

3. Enrichment Metric Analysis

Supposing a disease is associated with n genetic pathways, then we define the gene-wise enrichment score (EScore) for the *k*th gene within a gene set as given by Eq. (4).

$$EScore_k = \sum_{(i=1)}^{m} (-\log_{10} pValue_i) / \max_{(1< i< n)}(-\log_{10} pValue_i) \quad (4)$$

where, $pValue_i$ is the enrichment score of the ith pathway within the gene set; n is the total number of pathways; m the number of pathways including the *k*th gene.

4. Enrichment Analysis

To better understand the underlying functional profile and the pathogenic significance of the reported genes, we performed a gene set/pathway enrichment analysis (GSEA)and a sub-network enrichment analysis (SNEA) on 3 groups: 1) The whole gene list (235 genes); 2) Two subgroups selected using the highest quality matrix scores. In addition, we conducted a network connectivity analysis (NCA) using the Pathway Studio network building module.

RESULTS

1. Summary of LDM Results

In this study, we conducted a LDM of 2,410 articles that reported 235 genes associated with ADHD. According to the reported category of gene-ADHD relationship, the 2,410 articles can be clustered into 5 different classes: 1) Biomarker (0.21%); 2) Clinical Trial (0.04%); 3) Genetic Change (78.80%); 4) Quantitative Change (4.73%); and 5) Regulation (16.22%). Moreover, 26.81% genes have been reported to show multiple relationships with the disease.

For the 235 genes, 1.27% genes presented Biomarker relationship to the disease, 0.32% with Clinical Trial, 58.23% with Genetic Change, 12.03% with Quantitative Change, and 28.16% with Regulation. Moreover, 26.81% genes were reported to have multiple relationships with the disease. Specifically, 73.19% genes presented 1 type of relationship to the disease, 20.00% with 2, 5.96% with 3, and 0.85% with 4, as shown in Fig. 1.

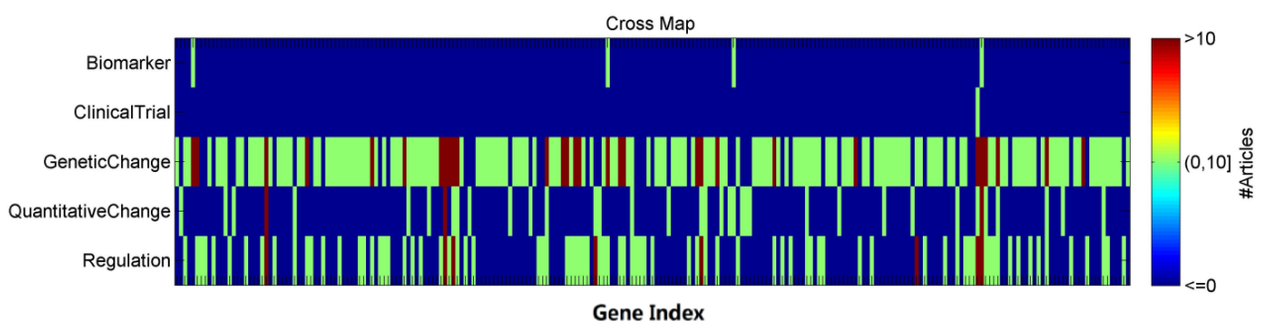


Fig.1 Gene-wise Relation Type Distribution of 235 Genes

The publication date distribution of these 2,410 articles is presented in Fig. 2 (a), where we show that this study covers literature data from the past 28 years (1988 - 2016), with novel genes reported in each year (Fig. 2 (b)). To note, these articles have an average publication age of only 6.4 years, indicating that most of the articles were published in recent years. In addition, recent years saw an increased number of publications, especially after 2010, with more novel genes being discovered (Fig.2 (b)). Moreover, our analysis showed that the publication date distributions of the articles underlying each of the 235 genes were also similar to that presented in Fig. 2.

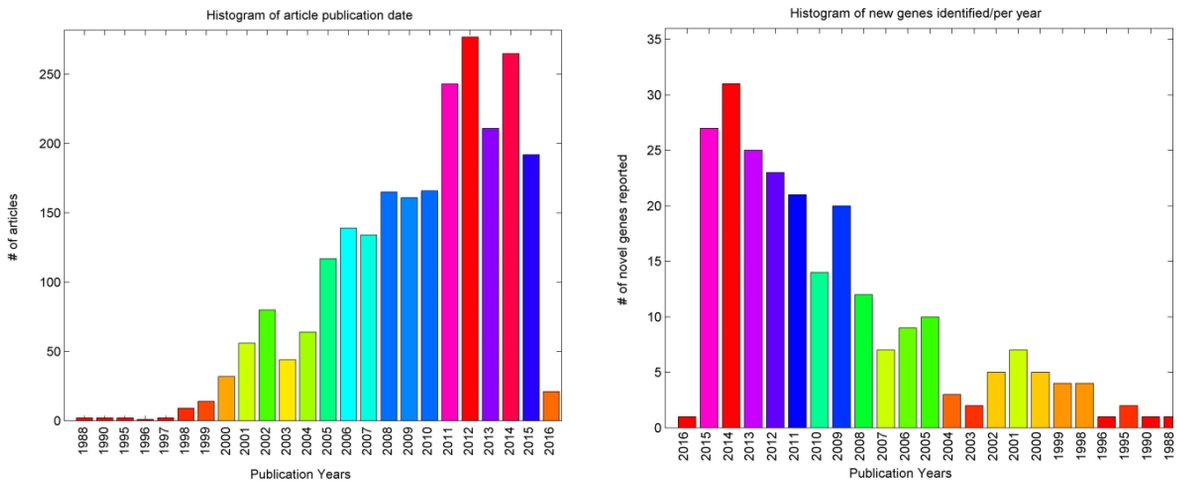


Fig. 2 Histogram of the Publications Reporting Gene-disease Relationships between ADHD and 235 Genes. (a) presents the histogram of article publication date; (b) presents the histogram of the number of novel genes identified in each year;

2. Marker Ranking

Using the 2 literature metrics scores, we identified that some genes were frequently reported with large numbers of articles to support them, such as SLC6A3 (366 articles), DRD4 (285 articles) and SLC6A2 (111 articles). These genes are the ones with highest RScores. Some genes have also been recently reported (e.g., reported within last two years) such as AS3MT, ANKK1 and MAP1B.

Among the 235 genes, 26 were reported within two years (2015-2016), which are listed in Table 1 and the full results are provided in **Supplementary Material 1.** For comparison purposes, Table 1 also lists the top 26 genes with the highest RScore.

Table 1 Top 26 Genes Reported Associations with ADHD Ranked by Different Scores

Genes with AScore ≤ 2	AS3MT;ANKK1;MAP1B;PER2;TRAF4;AIRE;CASP3;CTNNA2;F12;ITIH3;C10orf32;CHAF1B;DISC1;EP300;FGFR2;FKBP5;ID2;IDDM2;ZCCHC16;LGALS3;NOS2;PER1;SCN8A;SHANK2;STIP1;TNFRSF10A
Genes By RScore	SLC6A3;DRD4;SLC6A2;BDNF;COMT;SLC6A4;SNAP25;MAOA;PVRL2;DRD2;ADRA2A;HTR2A;DBH;LPHN3;MC4R;HTR1B;DRD5;GIT1;GRM7;MTHFR;ADRA2C;TPH2;DRD3;GRIN2A;NTF3;TH

3. Enrichment Analysis

In this section, we present the GSEA and SNEA results for 3 different groups: all 235 genes, and the 2 gene groups listed in Table 1.

3.1 Enrichment Analysis on All 235 Genes

The full list of 100 pathways/gene sets enriched with p-value<1.1e-007 (with 181 unique genes) is given in **Supplementary Material 2.** 51 pathways/gene sets were enriched with p-values<1E-10 (with 168 unique genes), 11 were enriched with p-values<1E-20 (with 139 unique genes), and 1 pathways/genes were enriched with p-values<1.7e-033 (47 unique genes). In Table 2, we present the top 20 pathways/groups enriched by all the genes, with p-values<2.6e-017 (155 unique genes).

Moreover, among these 100 pathways/gene sets enriched, we identified 15 pathways/gene sets (with 136 unique genes) that are related to the neuronal system, 9 pathways/gene sets (27 unique genes) were related to neuro transmitter, 7 pathways/gene sets (43 unique genes) were related to brain function/development and 5 pathways/gene sets (42 unique genes) were related to behavior. In Table 2, the Jaccard similarity (J_x), a statistics measure used for comparing the similarity and diversity of sample sets defined by Eq. (5), is given.

$$J_s(A,B) = \frac{A \cap B}{A \cup B} \qquad (5)$$

where, A and B are two sample sets.

Table 2 Molecular Function Pathways/Groups Enriched by 235 Genes Reported

Pathway/gene set name	Hit type	GO ID	# of Entities	Overlap	p-value	Jaccard similarity
synaptic transmission	biological_process	0007268	472	47	1.66E-33	0.07
dendrite	cellular_component	0030425	396	41	1.82E-30	0.07
neuron projection	cellular_component	0043005	378	40	4.29E-30	0.07
signal transduction	biological_process	0023033	1843	74	5.22E-27	0.04
synapse	cellular_component	0045202	466	40	1.43E-26	0.06
axon	cellular_component	0030424	318	33	1.31E-24	0.07
postsynaptic membrane	cellular_component	0045211	227	29	2.68E-24	0.07
neuronal cell body	cellular_component	0043025	466	37	2.08E-23	0.06
learning	biological_process	0007612	65	19	3.02E-23	0.07
memory	biological_process	0007613	76	19	8.54E-22	0.07
dendritic spine	cellular_component	0043197	135	22	5.01E-21	0.07
response to amphetamine	biological_process	0001975	41	15	2.66E-20	0.06
response to ethanol	biological_process	0017036	161	23	3.06E-20	0.06
social behavior	biological_process	0035176	52	16	3.66E-20	0.06
postsynaptic density	cellular_component	0014069	168	23	3.75E-20	0.06
response to cocaine	biological_process	0042220	44	15	9.38E-20	0.06
integral component of plasma membrane	cellular_component	0005887	1360	52	2.86E-18	0.03
response to drug	biological_process	0017035	509	33	8.77E-18	0.05
drug binding	molecular_function	0008144	110	19	9.77E-18	0.06
locomotory behavior	biological_process	0007626	108	18	2.58E-17	0.06

Besides the 8 neuronal system related pathways listed in Table 2, there are another 7 pathways/gene sets related to neuron system (P-value: [7.7e-015,6.6e-008]) and 9 ontology terms related to neuronal transmitter (P-value: [3.5e-014,1.1e-008]):neuronal postsynaptic density (GO: 0097481; p-value=7.7e-015, overlap: 14); signal transducer activity (GO: 0009369; p-value=4e-014, overlap: 36); synaptic transmission, dopaminergic (GO: 0001963; p-value=2.3e-013, overlap: 8); modulation of synaptic transmission (GO: 0050804; p-value=6.4e-009, overlap: 7); excitatory synapse (GO: 0060076; p-value=1.8e-008, overlap: 7); regulation of dopamine uptake involved in synaptic transmission (GO: 0051584; p-value=4.7e-008, overlap: 4); synapse assembly (GO: 0007416; p-value=6.6e-008, overlap: 8); dopamine metabolic process (GO: 0042417; p-value=3.5e-014, overlap: 9); dopamine binding (GO: 0035240; p-value=2e-012, overlap: 7); regulation of dopamine secretion (GO: 0014059; p-value=1.4e-011, overlap: 7); dopamine biosynthetic process (GO: 0042416; p-value=1.4e-011, overlap: 7); dopamine catabolic process (GO: 0042420; p-value=9.1e-011, overlap: 5); dopamine neurotransmitter receptor activity (GO: 0004952; p-value=1.5e-010, overlap: 5); adenylate cyclase-activating dopamine receptor signaling pathway (GO: 0007191; p-value=7.8e-010, overlap: 6); dopamine receptor signaling pathway (GO: 0007212; p-value=1e-008, overlap: 6); regulation of dopamine metabolic process (GO: 0042053; p-value=1.1e-008, overlap: 5).

Moreover, there were7additional pathways/gene sets related to memory and learning with relatively less significant p-values: learning (GO: 0007612; p-value=3e-023, overlap: 19); memory (GO: 0007613; p-value=8.5e-022, overlap: 19); associative learning (GO: 0008306; p-value=1.5e-012, overlap: 10); learning or memory (GO: 0007611; p-value=2.3e-011, overlap: 11); visual learning (GO: 0008542; p-value=4.6e-009, overlap: 9); long-term memory (GO: 0007616; p-value=2.9e-008, overlap: 7); cognition (GO: 0050890; p-value=1e-007, overlap: 7).

In addition, results showed 5 pathways/gene sets related to behavior (P-value: [3.7e-020,3.8e-009]): social behavior (GO: 0035176; p-value=3.7e-020, overlap: 16); locomotory behavior (GO: 0007626; p-value=2.6e-017, overlap: 18); behavioral response to cocaine (GO: 0048148; p-value=1e-013, overlap: 8); adult locomotory behavior (GO: 0008344; p-value=4.6e-011, overlap: 11); behavioral fear response (GO: 0001662; p-value=3.8e-009, overlap: 8).

Besides GSEA, we also performed a SNEA using Pathway Studio with the purpose of identifying the pathogenic significance of the reported genes to other disorders that are potentially related to ADHD. We provide the full list of results in **Supplementary Material 3.** In Table 3, we present the disease related sub-networks enriched with a p-value<7.05E-46.

Table 3 Sub-networks Enriched by the 235 Genes Reported

Gene Set Seed	Total # of Neighbors	Overlap	p-value	Jaccard similarity
Attention Deficit Disorder with Hyperactivity	358	81	6.7E-119	0.17
Schizophrenia	1287	85	1.83E-77	0.06
Mental Disorders	721	69	4.98E-73	0.08
Anxiety	1621	79	2.71E-61	0.05
Aggression	1472	72	1.55E-55	0.04
cognitive impairment	1542	69	1.71E-50	0.04
Cognition Disorders	1000	57	4.85E-47	0.05
Behavior, Addictive	412	43	1.63E-46	0.07
neuropathic pain	1337	62	5.49E-46	0.04
motor dysfunction	1221	60	7.05E-46	0.04

From Table 3, we see that many of these reported ADHD related genes were also identified in other mental health related diseases, with a large percentage of overlap (Jaccard similarity ≥ 0.04).

3.2 Enrichment Analysis on Top 26 Genes with Highest Scores

We compared the top 26 genes listed in Table 1 in terms of GSEA and SNEA results. Here we only present the top 10 pathways/sub-networks for the AScore and the RScore groups respectively (Table 4 and Table 5), and report the full in Supplementary Material 2 and 3.

Using the same enrichment p-value threshold (p<1E-3), we identified 41 pathways/gene sets that were enriched with the 26 genes with top AScores, while the number of genes for RScore group is 181. The full lists of these pathways/gene sets are provided in Supplementary Material 2. Table 4 presents the top 10 pathways enriched with the 26 genes from AScore and RScore groups, respectively.

Table 4 Pathways/groups Enriched by 26 Genes with the Highest AScore and RScore

	Pathway/gene set Name	GO ID	p-value

The first 10 pathways/ gene sets enriched by top 26 genes with highest AScores	positive regulation of axon extension	0045773;	3.7E-06
	circadian regulation of translation	0097167;	6.88E-06
	endodermal digestive tract morphogenesis	0061031;	6.88E-06
	circadian rhythm	0007623;	9.73E-06
	response to tumor necrosis factor	0034612;	1.02E-05
	PAC (PAS-associated C-terminal) domain	Pathway Studio Ontology	1.46E-05
	membranous septum morphogenesis	0003149;	1.72E-05
	regulation of circadian rhythm	0042752;	2.91E-05
	PAS (PER-ARNT-SIM) domain	Pathway Studio Ontology	3.05E-05
	entrainment of circadian clock	0009649;	4.12E-05
The first 10 pathways/ gene sets enriched by top 26 genes with highest RScore	axon	0030424;	5.84E-16
	dopamine binding	0035240;	7.15E-16
	synaptic transmission	0007268;	2.55E-15
	response to drug	0017035;	6.73E-15
	response to amphetamine	0001975;	4.05E-14
	locomotory behavior	0007626;	5.46E-13
	synaptic transmission, dopaminergic	0001963;	3.85E-12
	dopamine catabolic process	0042420;	9.24E-12
	response to cocaine	0042220;	1.33E-11
	dopamine neurotransmitter receptor activity	0004952;	1.35E-11

From Table 4, we see that the genes with the top AScores and those with the top RScores were enriching different groups of pathways, with different p-values (AScore group: 3.7E-06~4.12E-05; RScore group: 5.84E-16~1.35E-11), indicating that the newly reported genes are functionally different from the most frequently reported ones. Moreover, we observed that 6 out of the 10 pathways/gene sets enriched by the RScore group (Table 4) were observed in Table 2, which lists the top 20 pathways/gene sets enriched with 155 /235 genes, whereas the number for AScore group is 0.

For the SNEA analysis, we only performed an enrichment analysis against disease sub-networks. We provide the full list of results in Supplementary Material 3. Table 5 presents the top 10 disease related sub-networks enriched by the top 26 genes from AScore group and RScore group, respectively.

Table 5 SNEA Results by 26 Genes with the Highest AScore and RScore

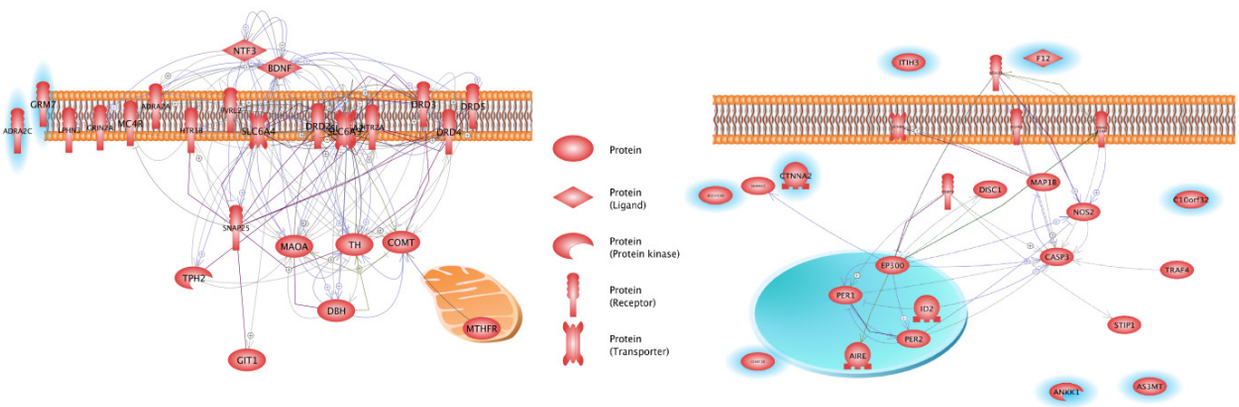
	Gene Set Seed	Overlap	p-value	Jaccard similarity
The first 10 pathways/ gene sets enriched by top 26 genes with highest AScores	Bipolar Disorder	10	9.08E-12	0.02
	Schizophrenia	13	9.39E-12	0.01
	Anxiety	8	1.29E-09	0.02
	Psychotic Disorders	7	8.25E-09	0.02
	Cancer of Head and Neck	7	1.49E-08	0.02
	Diabetes Mellitus	13	3.07E-08	0.01
	Obesity	12	3.55E-08	0.01
	Autism Spectrum Disorders	7	1.04E-07	0.02
	Alcoholism	7	1.08E-07	0.02
	Prenatal Exposure Delayed Effects	5	1.34E-07	0.03

The first 10 pathways/ gene sets enriched by top 26 genes with highest RScore	Attention Deficit Disorder	14	1.12E-33	0.26
	Depressive Disorder, Major	21	9.59E-33	0.05
	Neurotic Disorders	13	6.06E-30	0.22
	Bipolar Disorder	20	1.69E-29	0.04
	Behavior, Addictive	15	7.19E-29	0.11
	Alcoholism	19	2.60E-28	0.04
	Obsessive-Compulsive Disorder	13	1.49E-27	0.16
	Stress Disorders, Post-Traumatic	14	2.13E-27	0.12
	Personality Disorders	11	2.21E-27	0.26
	Tourette Syndrome	13	2.79E-27	0.16

From Table 5, we see that both groups enriched other mental health related sub-networks. However, the enrichment p-values of the RScore group were much more significant than those shown by the AScore group, and with higher Jaccard similarities.

4. Connectivity Analysis

In addition to GSEA and SNEA, we performed a NCA on the top 26 geneswith the highest RScores and AScores (from Table 1) to generate gene-gene interaction network. Results showed that for the RScore group, there were 154 connections among 24/26 genes, which are strongly supported by the literature. Only 2 genes present no direct connection with any other genes (Fig. 3 (a); highlighted in blue). In contrast, genes within the AScore group demonstrated only 36 relations among 18/26 genes, as shown in Fig. 3 (b), with 8 genes showing no direct relation with other genes in the group (Fig. 3 (b); highlighted in blue). This observation is consistent with the GSEA and SNEA, suggesting that genes with the smallest AScores are not as functionally close to each other as those from the RScore group.

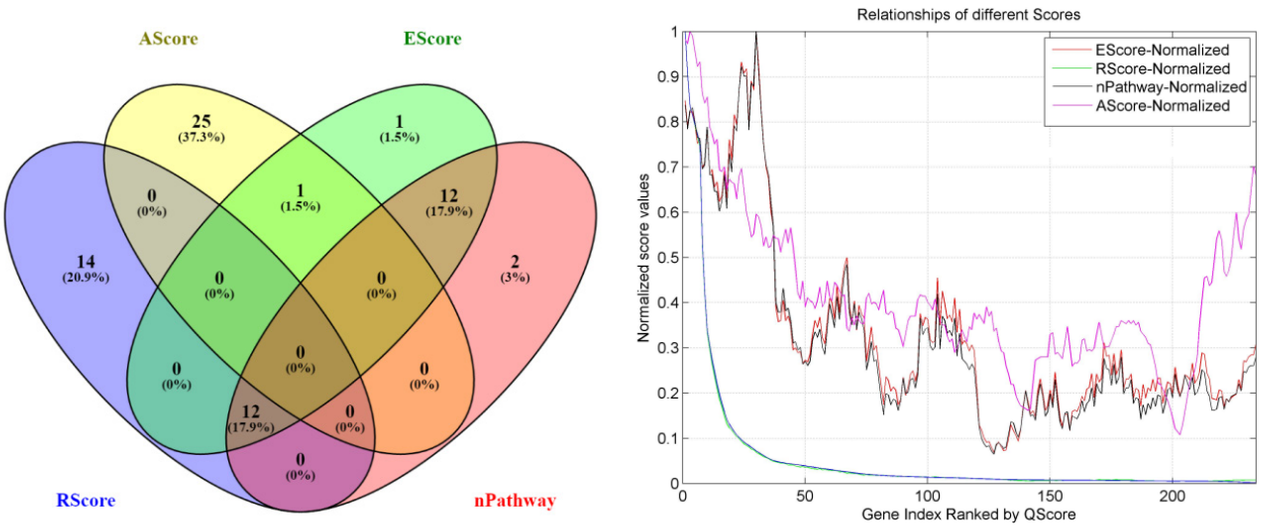


(a) By RScore group (b) By AScore group

Fig. 3 Connectivity Networks Built by 26 Genes from Different Groups. The Networks Were Generated Using Pathway Studio; The Un-related Genes Are Highlighted in Blue.

5. EScore Analysis

Through GSEA, we also generated a biological metrics, EScore, for each gene. The value of an ESocre represents how a gene is related to the pathways associated with ADHD. To compare the EScore and the literature metrics, we performed a cross-analysis of the top 26 genes selected using different scores, and present a Venn diagram in Fig.4 (a) (Oliveros, 2007-2015). EScore and nPathway groups showed an overlap of 24/26 (with the exception of MECP2 and NOS1). EScore group has a relatively large overlap with RScore group (overlap 12/26):SLC6A3; DRD4; BDNF; DRD2; HTR2A; GRM7; DBH; HTR1B; DRD5; TH; DRD3, GRIN2A, whereas an overlap of only one gene (SHANK2) with AScore group. To note, there was one gene, GRM5, that has been exclusively included in EScore, showing RScore of only 7.



(a) (b)

Fig.4 Comparison of Different Metrics Ranking the 235 Genes. (a) A Venn diagram of top 26 genes selected by different metrics; (b) Comparison of average metrics values with gene set size of 7

Besides comparing the top 26 genes (Fig. 4 (a)), we also compared the averaged metric values of all the 235 genes on a group level, as shown in Fig.4 (b). We used a group size of 7 genes, that is, we first sorted the 235 genes by RScore, then we averaged each type of metrics value using a moving window of length 7. Results showed that the average scores were strongly correlated, as shown in Table 6.

Table 6 Pearson Correlation Coefficients between Different Metrics

	RScore	EScore	nPathway	AScore
RScore	1.00	0.60	0.61	0.74
EScore	0.60	1.00	1.00	0.75
nPathway	0.61	1.00	1.00	0.76
AScore	0.74	0.75	0.76	1.00

DISCUSSION

In this work, we performed a LDM on 2,410 articles (from 1988 to April 2016), reporting 235 genes associated with ADHD. We provide in Supplementary Materials 1 the full gene list together with the literature and enrichment metrics scores. In addition, results from GSEA and SNEA support the current literature that most of these genes may play roles in the pathogenesis of ADHD. Furthermore, NCA showed that many of these genes were functionally linked to one another.

As an automatic data mining approach, the NLP technique is effective and efficient in dealing with large amounts of literature data for LDM. However, the LDM method may produce some false positives. Therefore, the results of this study are intended to provide an overview map for the current field of genetic studies of ADHD and lay the groundwork for further biological/genetic studies in this area.

Although our analysis did not specifically focus on single genes, we noticed that the 235 genes identified were not equal in terms of publication frequency (RScore), their novelties (AScore) and the functional diversity (EScore). Using the proposed quality metrics scores, one is able to rank the genes according to

different needs/significance and pick the top ones for further analysis (Table 1). For example, the top 5 genes by AScore, namely AS3MT, ANKK1, MAP1B, PER2and TRAF4, are the ones that were recently reported. On the other hand, SLC6A3, DRD4, SLC6A2, COMT and BDNF are the top 5 genes that were found to be most often replicated in studies (with highest RScores), suggesting them as common variables in the occurrence of ADHD. These genes likely possess biological significance in relation to the disease.

Additionally, we noted that, for the top 100 pathways enriched with 181/235 genes (Supplementary Material 2), some genes were duplicated in multiple significantly enriched pathways, presenting high EScore, such as DRD1 (44/100 pathways), DRD2 (42/100), GRIN1 (38/100), GRIN2A (35/100) and GRIN2B (35/100). These genes play multiple roles within different genetic pathways associated with ADHD, indicating the biological significance of these genes in association with the disease.

There is a relatively large overlap between the top ESocre and RScore genes (Fig.4), indicating that the frequently replicated genes tend to play roles within multiple pathways associated with ADHD. Moreover, one gene, SHANK2, reported recently in only a few article, demonstrated high EScore. SHANK2 is been included in 13 of the top 100 pathways, many of which have been implicated to be related to ADHD. These include neuron projection (0043005); synapse (0045202); postsynaptic membrane (0045211); neuronal cell body (0043025); learning (0007612); memory (0007613); dendritic spine (0043197); social behavior (0035176); postsynaptic density (0014069); long-term synaptic potentiation (0060291); adult behavior (0030534).(16,17) The observation suggested that SHANK2 is worthy of further study with regard to the pathogenesis of ADHD.

Additionally, we observed that most genes identified by this LDM were included in the pathways previously implicated with ADHD, including 15 neuronal system pathways, 9 neuronal transmitter pathways, 7 brain function related pathways and 5 behavior related gene ontology terms.(16,19) To note, 181/235 were included in the top 100 enriched pathways (p-value< 1.1e-007), and 155/235 in the top 20 pathways listed in Table 2 (p-value< 2.6e-017). We hypothesize that the majority of these literature reported genes, especially the ones that were identified from significantly enriched pathways, should be functionally linked to ADHD. Although there may be false positives from the separate studies undertaken in the different literature publications, it is less likely that a large group of genes were falsely perturbed at the same time than a single gene was, which is one of the advantages of GSEA. (15)

Another advantage of GSEA is that, when the members of a gene set exhibit strong cross-correlation, GSEA can boost the signal-to-noise ratio and make it possible to detect modest changes in individual genes. (15) The NCA analysis showed that many of the 235 reported genes were functionally associated with one another (Fig.3), indicating that these functionally related genes from literature possess higher probabilities as true hits than that as noise (false positives).

In addition to GSEA, we performed a sub-network enrichment analysis (SNEA), which was implemented in Pathway Studio using master casual networks, a database containing more than 6.5 million relationships derived from more than 4 million full text articles and 25 million PubMed abstracts. These networks were generated by a finely-tuned NLP text mining system to extract relationship data from the scientific literature. The ability to quickly update the terminologies and linguistic rules used by NLP systems ensures that new terms can be captured soon after entering into regular use in the literature. This extensive database of interaction data provides high levels of confidence when interpreting experimentally-derived genetic data against the background of previously published results (http://help.pathwaystudio.com/fileadmin/standalone/pathway_studio/help_ps_10.0/index.html?analyze_experiment.htm). Here, SNEA results demonstrated that many of the 235 genes (>90%) showing strong association with ADHD were also identified as causal genes involved in other mental health disorders (schizophrenia, aggression, cognitive impairment).(20-22)

Nevertheless, this study has some limitations that should be considered in future work. The literature data of the 2,410 articles studied were extracted from the Pathway Studio database. Although Pathway Studio database covers over 40 million articles, it is still possible that some articles studying gene-ADHD associations were beyond their scope of coverage. Additionally, the quality scores, RScore, AScore, and EScore were proposed as quality measures of the literature reported gene-disease relations. Although related to, they are not the direct biological significance measures of relationship of the genes to the disease.

CONCLUSION

Results from this up-to-date LDM reveal that the 235 genes identified have multiple types of associations with ADHD, providing an overview map for the current genetic study of ADHD. Meanwhile, the literature and enrichment metrics discovered top genes with specific significance in relation to the disease. In addition,

NCA and enrichment analysis results suggested that these genes play significant roles as a network in the pathogenesis of ADHD, operating as a functional genetic network influencing the development of ADHD. At the same time, the same genes may also play an important role in the pathogenesis of many other ADHD related disorders.

We conclude that ADHD is a complex disease for which genetic causes are linked to a network composed of a large group of genes. LDM together with GSEA, SNEA and NCA could serve as an effective approach in finding these potential target genes. This study provided an overview map, with different metrics, for the current field of genetic studies of ADHD, which could be used as a groundwork for further biological and genetic studies in the field.

Declaration of Interests

The authors declare no conflict of interests.

SUPPLEMENTARY

1. Supplementary material 1 at http://www.qingres.com/Upload/Excel/Supplementary material 1.xlsx
2. Supplementary material 2 at http://www.qingres.com/Upload/Excel/Supplementary material 2.xlsx
3. Supplementary material 3 at http://www.qingres.com/Upload/Excel/Supplementary material 3.xlsx

REFERENCES

1. Sroubek A, Kelly M, Xiaobo Li. Inattentiveness in attention-deficit/hyperactivity disorder. Neuroscience Bulletin. 2013; 29 (1): 103–10.
2. Global Burden of Disease Study 2013, Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet; 2015; 386(9995): p743–800.
3. Singh, I. Beyond polemics: Science and ethics of ADHD. Nature Reviews Neuroscience. 2008; 9(12): 957–64.
4. Thapar A, Cooper M, Eyre O, Langley K. What have we learnt about the causes of ADHD? J Child Psychol Psychiatry. 2013; 54 (1): 3-16.
5. Giana G, Romano E, Porfirio MC, D'Ambrosio R, Giovinazzo S, Troianiello M, et al. Detection of auto-antibodies to DAT in the serum: Interactions with DAT genotype and psycho-stimulant therapy for ADHD. Journal of Neuroimmunology. 2014; 278:212-222.
6. Cummins TDR, Hawi Z, Bellgrove MA, Cummins TDR, Jacoby O, Hawi Z, et al. Alpha-2A adrenergic receptor gene variants are associated with increased intra-individual variability in response time. Molecular Psychiatry. 2014; 19(9):1031-1036.
7. Kim BN, Kim JW, Hong SB, Cho SC, Shin MS and Yoo HJe. Possible association of norepinephrine transporter-3081(A/T) polymorphism with methylphenidate response in attention deficit hyperactivity disorder. Behavioral and Brain Functions. 2010; 6:57.
8. Kovtun O, Sakrikar D, Tomlinson ID, Chang JC, Arzeta-Ferrer X, Blakely RD, et al. Single-quantum-dot tracking reveals altered membrane dynamics of an attention-deficit/hyperactivity-disorder-derived dopamine transporter coding variant. ACS Chemical Neuroscience. 2015; 6(4):526-534.
9. Van Mil NH, Steegers-Theunissen RPM, Bouwland-Both MI, Verbiest MMPJ, Rijlaarsdam J, Hofman A, et al. DNA methylation profiles at birth and child ADHD symptoms. Journal of Psychiatric Research. 2013; 49(1):51-59.
10. Frank MK, de Mello MT, Lee KS, Daubian-Nose P, Tufik S, Esteves AM. Sleep-related movement disorder symptoms in SHR are attenuated by physical exercise and an angiotensin-converting enzyme inhibitor. Physiol Behav. 2016; 154:161-8.
11. Hawi Z, Matthews N, Wagner J, Wallace RH, Butler TJ, Vance A, et al. DNA Variation in the SNAP25 Gene Confers Risk to ADHD and Is Associated with Reduced Expression in Prefrontal Cortex. PLoS ONE. 2013;8(4): e60274.

12. Ma Y, Krueger JJ, Redmon SN, Uppuganti S, Nyman JS, Hahn MK, et al. Extracellular norepinephrine clearance by the norepinephrine transporter is required for skeletal homeostasis. Journal of Biological Chemistry. 2013; 288(42):30105-30113.

13. Nyman ES, Loukola A, Varilo T, Taanila A, Hurtig T, Moilanen I, et al. Sex-specific influence of DRD2 on ADHD-type temperament in a large population-based birth cohort. Psychiatric Genetics. 2012; 22(4):197-201.

14. Hong Q, Wang YP, Zhang M, Pan XQ, Guo M, Li F, et al. Homer expression in the hippocampus of an animal model of attention-deficit/hyperactivity disorder. Molecular Medicine Reports. 2011; 4(4):705-712.

15. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102(43):15545-50.

16. Tsai SJ. Signal transducer and activator of transcription 6 (STAT6) and attention-deficit hyperactivity disorder: a speculative hypothesis. Med Hypotheses. 2006; 67(6):1342-4.

17. Alderson RM, Rapport MD, Kofler MJ. Attention-Deficit/Hyperactivity Disorder and Behavioral Inhibition: A Meta-Analytic Review of the Stop-signal Paradigm. Journal of abnormal child Psychology. 2007; 35(5):745-758.

18. Dougherty DD, Bonab AA, Spencer TJ, Rauch SL, Madras BK, Fischman AJ. Dopamine transporter density in patients with attention deficit hyperactivity disorder. The Lancet. 1999; 354(9196): 2132-33.

19. Cantwell DP, Baker L. Association between attention deficit-hyperactivity disorder and learning disorders. J Learn Disabil. 1991;24(2):88-95.

20. Pallanti S, Salerno L. Raising attention to attention deficit hyperactivity disorder in schizophrenia. World J Psychiatry. 2015; 5(1): 47–55.

21. King S, Waschbusch DA. Aggression in children with attention-deficit/hyperactivity disorder. Expert Rev Neurother. 2010; 10(10):1581-94.

22. Kuntsi J, Wood AC, Rijsdijk F, Johnson KA, Andreou P, Albrecht B, et al. Separation of cognitive impairments in attention-deficit/hyperactivity disorder into 2 familial factors. Arch Gen Psychiatry. 2010; 67(11):1159-67.