Saudi Journal of Biological Sciences xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Saudi Journal of Biological Sciences

journal homepage: www.sciencedirect.com

Original article

Integrative computational evaluation of genetic markers for Alzheimer's disease

Zhe Li [a,b,1], ZhenZhen Xiong [c,1], Lydia C. Manor [d], Hongbao Cao [e,f], Tao Li [a,*]

[a] The Mental Health Center and the Psychiatric Laboratory, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China
[b] The Mental Rehabilitation Center, Karamay Municipal People's Hospital, Karamay, Xinjiang 830054, China
[c] School of Nursing, Chengdu Medical College, Chengdu, Sichuan 610083, China
[d] Department of Bioinformatics, American Informatics Consultant LLC, Rockville, MD 20852, USA
[e] Unit on Statistical Genomics, National Institute of Mental Health, NIH, Bethesda 20892, USA
[f] Department of Genomics Research, R&D Solutions, Elsevier Inc., Rockville, MD 20852, USA

ARTICLE INFO

Article history:
Received 1 May 2018
Revised 14 May 2018
Accepted 14 May 2018
Available online xxxx

Keywords:
Alzheimer's disease
ResNet database
Pathway enrichment analysis
Sub-network enrichment analysis
Gene-gene interaction analysis

ABSTRACT

Recent studies have reported hundreds of genes linked to Alzheimer's Disease (AD). However, many of these candidate genes may be not identified in different studies when analyses were replicated. Moreover, results could be controversial. Here, we proposed a computational workflow to curate and evaluate AD related genes. The method integrates large scale literature knowledge data and gene expression data that were acquired from postmortem human brain regions (AD case/control: 31/32 and 22/8). Pathway Enrichment, Sub-Network Enrichment, and Gene-Gene Interaction analysis were conducted to study the pathogenic profile of the candidate genes, with 4 metrics proposed and validated for each gene. By using our approach, a scalable AD genetic database was developed, including AD related genes, pathways, diseases and info of supporting references. The AD case/control classification supported the effectiveness of the 4 proposed metrics, which successfully identified 21 well-studied AD genes (i.g. TGFB1, CTNNB1, APP, IL1B, PSEN1, PTGS2, IL6, VEGFA, SOD1, AKT1, CDK5, TNF, GSK3B, TP53, CCL2, BDNF, NGF, IGF1, SIRT1, AGER and TLR) and highlighted one recently reported AD gene (i.g. ITGB1). The computational biology approach and the AD database developed in this study provide a valuable resource which may facilitate the understanding of the AD genetic profile.

1. Introduction

Alzheimer's disease (AD) is a chronic neurodegenerative disease that usually onsets slowly and progresses more rapidly over time (Burns and Iliffe, 2009). It is the leading cause of dementia, beginning with impaired memory, and most often onsets in people over 65 years of age (Mendez 2012). The global prevalence of AD as of 2015 was estimated to be as high as 48 million people worldwide (World Health Organization, 2015). Although the cause of most Alzheimer's cases largely remains unknown, about 70% of the risk is believed to come from a large network of genes (Ballard et al., 2011). As such, researches into the causes of AD are currently being explored.

In recent years, an increased number of genetic researches have been conducted revealing over a thousand altered genes linked to AD. For example, increased GSK3B activity and decreased phosphorylation of the gene have been repeatedly observed in AD cases (Cole et al., 2007; Koedam et al., 2013; Xu et al., 2016). Significantly increased expression levels of TP53, PTGS2 and TGFB1 were suggested by many independent studies to be associated with AD (Cenini et al., 2008; Lanni et al., 2012; Ramalho et al., 2008; Yoo et al., 2008; Wang et al., 2014; Luo et al., 2006). Observations from these previous studies are valuable in studying the genetic basis of the pathogenic development of the disease.

However, approximately one third of these AD-gene linkages were reported once with no further replication, and over 60% were supported by no more than three citations. Moreover, most of

* Corresponding author at: The Mental Health Center and the Psychiatric Laboratory, West China Hospital, Sichuan University, No. 28 Dian Xin Nan Road, Chengdu, Sichuan 610041, China.
E-mail address: xuntao26@hotmail.com (T. Li).
[1] These authors contributed equally to this work.

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.sjbs.2018.05.019
1319-562X/

these studies had small sample sizes that were more susceptible to noise. Additionally, due to the variation in data collection and processing approaches, results from different studies were not always consistent. Meanwhile, there are dozens of new AD risk genes being reported every year, posing an increased need for further validation of these candidate genes to AD. While biological experiments were effective towards this validation task, they could be very costly. To address this issue, we propose a computational biology approach for a systematic evaluation of these AD candidate genes.

In recent years, Pathway Studio ResNet relation data have been widely used to study modeled relationships between proteins, genes, complexes, cells, tissues and diseases (http://pathwaystudio.gousinfo.com/Mendeley.html). In this study, we integrated large scale AD related ResNet literature knowledge data, independent gene expression data and related pathway/network information to study the functional profile of a large gene pool that has been reported to be linked to AD. The purpose of the study is to provide an easy-update computational evaluation workflow, through which an AD genetic database (AD_GD) could be generated to present a weighted landscape view of the genetic basis underlying the pathogenic development of AD. Our results support the hypothesis that AD candidate genes are functionally linked to each other, forming a large genetic network to regulate the pathogenic development of AD through multiple pathways.

2. Materials and methods

Fig. 1 presents the diagram of the proposed computational gene marker evaluation system, with detailed descriptions in the following sub-sections. Using our approach, a genetic database (AD_GD) was developed and deposited into an open source 'Bioinformatics Database' online available at http://database.gousinfo.com, including 1699 genes (with metric scores), 151 pathways and 114 diseases that are linked to AD. Also included in AD_GD are information of 27000+ supporting references for AD-gene relationships, including the titles and relevant sentences where the relations were identified. The AD_GD database is scalable and will be updated monthly or upon request using our approach.

2.1. ResNet literature knowledge data

ResNet relation data (AD-Gene) were acquired from the Pathway Studio ResNet® Mammalian database (http://pathwaystudio.gousinfo.com/ResNetDatabase.html) updated November 2016. The ResNet® Mammalian database are a group of real-time updated literature knowledge databases, including curated signaling, cellular processes and metabolic pathways, ontologies and annotations, as well as molecular interactions and functional relationships (http://pathwaystudio.gousinfo.com/ResNetDatabase.html). Modeled relation data are extracted from the 41M+ references covering entire PubMed abstracts and Elsevier and third party full text journals. The ResNet database employs an automated natural language processing-based information extraction system, MedScan, with precision of over 91% (Daraselia et al., 2004). Each relationship within the database is supported with one or more references. By far, Pathway Studio ResNet Databases is the largest database among known competitors in the field (Lorenzi et al., 2014).

2.2. Enrichment and gene-gene interaction analysis

Pathway enrichment analysis (PEA) and sub-network enrichment analysis (SNEA) (http://pathwaystudio.gousinfo.com/SNEA.pdf) was conducted using Pathway Studio to identify genetic pathways and diseases potentially linked to AD (Sivachenko et al., 2007). Furthermore, a pathway based gene-gene interaction (GGI) analysis was conducted to generate weighted edges/linkage between genes. The weight of an edge is the number of pathways where both nodes were included.

2.3. Metrics analysis

For the gene network built through the aforementioned steps, 4 metrics were proposed for each node/gene, including 2 literature based metric scores (RScore and AScore), and 2 enrichment based metric scores (PScore and SScore). The logic is that, a gene is likely linked to AD if it satisfies one or more of the following conditions: the gene has been frequently observed in independent studies to be associated with AD (high RScore), plays roles within multiple pathways associated with AD (high PScore), and demonstrates strong functional linkage to many of other genes associated with AD (high SSCore). Additionally, an AScore was proposed to present the history of each AD-gene relation. The detailed definitions of the proposed metrics are described as follows.

2.3.1. Two literature metrics

The reference score (RScore) of a gene is defined as the reference number underlying a gene-disease relationship, as shown in Eq. (1).

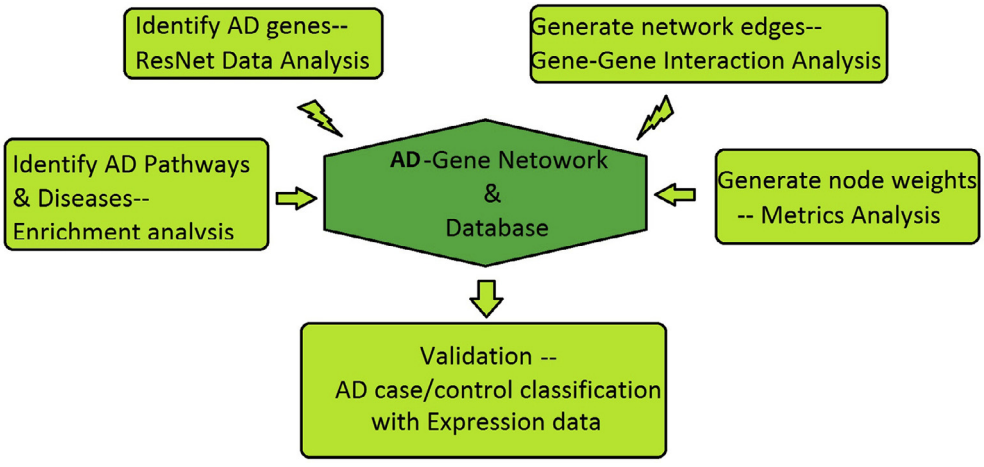

Fig. 1. Diagram for the integrative computational marker evaluation approach for AD. First, literature based analysis were conducted to identify the AD related genes, then Gene-Gene Interaction Analysis, Enrichment analysis, and Metrics analysis were conducted on these gene and results were saved in the AD database. Finally, AD case/control classification were conducted to test the effectiveness of the identified genes, using gene expression datasets.

Please cite this article in press as: Li, Z., et al. Integrative computational evaluation of genetic markers for Alzheimer's disease. Saudi Journal of Biological Sciences (2018), https://doi.org/10.1016/j.sjbs.2018.05.019