Bridging Heterogeneous Mutation Data to Enhance Disease-Gene Discovery

Kaiyin Zhou
Huazhong Agriculture University
Yuxing Wang
Huazhong Agriculture University
Kevin Bretonnel Cohen
University of Colorado Denver - Anschutz Medical Campus
Jin-Dong Kim
Research Organization of Information and Systems
Xiaohang Ma
Huazhong Agriculture University
Zhixue Shen
Huazhong Agriculture University
Xiangyu Meng
Wuhan University Zhongnan Hospital
Jingbo Xia (xiajingbo.math@gmail.com)
Huazhong Agriculture University https://orcid.org/0000-0002-7285-588X

Research article

Keywords: Heterogeneous data, data fusion, generative model, GWAS, text mining

Posted Date: August 14th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-44127/v1

Version of Record: A version of this preprint was published at Briefings in Bioinformatics on April 13th, 2021. See the published version at https://doi.org/10.1093/bib/bbab079.

Bridging Heterogeneous Mutation Data to Enhance Disease-Gene Discovery

Kaiyin Zhou[1#] zhoukaiyinhzau@gmail.com, Yuxing Wang[1#] wang-yuxing@foxmail.com, Kevin Bretonnel Cohen[2] kevin.cohen@gmail.com, Jin-Dong Kim[3] jdkim@dbcls.rois.ac.jp, Xiaohang Ma[1] mxh.hzau.edu.cn@webmail.hzau.edu.cn, Zhixue Shen[1] zhixue_shen@163.com, Xiangyu Meng[4,5] mengxy_whu@163.com, Jingbo Xia[1*] xiajingbo.math@gmail.com

[1]Hubei Key Lab of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, Hubei Province, P.R. China
[2]School of Medicine, University of Colorado at Denver, Anschutz Medical Campus, Colorado, U.S
[3]Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS), Tokyo, Japan
[4]Department of Urology, Zhongnan Hospital of Wuhan University, Wuhan, Hubei Province, P.R. China
[5]Institut Curie, CNRS, Molecular Oncology Team, PSL Research University, Paris, France
[#]The authors have the same contributions
[*]Correspondence: xiajingbo.math@gmail.com, xjb@mail.hzau.edu.cn

Abstract

Background: Bridging heterogeneous mutation data fills in the gap between various data categories and propels discovery of disease-related genes. It is known that genome-wide association study (GWAS) infers significant *mutation associations* which link genotype and phenotype, and it is under-powered for pinpointing causal genes due to high false positive or negative rate. In the meantime, mutation events widely reported in literature unveil typical functional biological process, including *mutation types* like gain-of-function and loss-of-function.

Methods: To bring together the heterogeneous mutation data, we propose a pipeline, "Gene-Disease Association prediction by Mutation Data Bridging (GDAMDB)", with a statistic generative model. The model learns the distribution parameters of *mutation associations* and *mutation types*, and recovers false negative GWAS mutations which fail to pass significant test but represent supportive evidences of functional biological process in literature.

Results: Eventually, GDAMDB is applied in Alzheimer's disease which is a common inheritable neurodegenerative disorder with unknown pathological mechanism, and it predicted 79 AD-associated genes. Besides 12 of them come from the original GWAS study, 57 of them are supported to be AD-related by other GWAS or literature report.

Conclusion: Our model is capable of enhancing the GWAS-based gene association discovery by well combining text mining results. The positive result indicates that bridging the heterogeneous mutation data is contributory for the novel disease-related gene discovery.

Key Words: Heterogeneous data, data fusion, generative model, GWAS, text mining

1 Background

Genome-wide association study (GWAS) is helpful to identify disease-related genes through significant test on single nucleotide polymorphisms (SNPs) across the entire genome, and the p-value of the SNPs is the *mutation association* data that represent the relevance between the gene and disease. However, as generally recognized, the high false negative rate of GWAS makes that not all *de facto* vital variations are able to pass the multiple testing, and the lack of consideration about the biological mechanism between genes and phenotype makes GWAS insufficient to pinpoint causal variations. Therefore, various researches considered to combine other Omics data with GWAS. For example, Jia et al. [1] developed a dense module searching method, which combined protein-protein interaction (PPI) network with GWAS data, to identify the candidate genes that was well studied in PPI but with low *p*-value in GWAS [2]. Wang et al. [3] proposed that genes in the same functional pathway may work together to raise a disease but some of them are difficult to reach the significant threshold in GWAS. Therefore, they combined the pathway knowledge with existed GWAS data to retrieve the omitted genes. Though it has been attempted to combine various Omics data with GWAS data, to combine another type of heterogeneous mutation data with *mutation association* GWAS is still a new idea.

The *mutation type*, i.e., loss of function (LOF) or gain of function (GOF) [4], is the mutation categorical data which bridges genes and diseases [5]. The LOF mutation in a gene results in the reduction or abolition of the gene function, while GOF mutation in a gene results in the enhanced or new gene function. The changed genes function eventually leads to the downstream molecular and cellular biological process, well supports the investigation of gene-disease associations, and provides evidence to unveil the pathological mechanism of diseases.

To our best knowledge, there is no disease-related LOF/GOF database. Fortunately, the relevant biological processes representing LOF or GOF are abundant in literature. Taking Alzheimer's disease (AD) as an example, *mutation types* reported in literature represent supportive information in functional biological process. In 1997, Citron et al. [6] proposed that the GOF mutation on PSEN1 and PSEN2 lead to an increase in $A\beta_{42}$ production. In 2013,

Guerreiro et al. [7] analyzed the genetic variability of TREM2 in 1,092 AD patient and 1,107 controls and found the LOF mutation in TREM2 are associated with AD. In 2015, Steinberg et al. [8] identified that LOF variants in ABCA7 are related with AD in Icelanders. Though the *mutation type* information is usually well described in literature, most of them are stated in an explicit semantics. For instance, the sentence *"In studies of cell lines transfected with beta-amyloid precursor protein (beta APP) cDNAs, the beta APP mutation K670N/M671L found in a Swedish familial AD (FAD) pedigree has previously been shown to cause a marked augmentation of A beta secretion."(pmid:7991571)* described a GOF *mutation type*, while *"In this issue of Neuron, describe two rare ADAM10 prodomain mutations that cause late-onset Alzheimer's disease by impairing prodomain chaperone function, attenuating alpha-secretase activity, and reducing adult hippocampal neurogenesis."(pmid:24139026)* described a LOF mutation type.

Owing to the rapid growth of algorithm strength, a combination of the state-of-the-art text mining strategy and the customized corpus make it possible to extract *mutation type* from literature mining in a PubMed scale. An active gene annotation corpus (AGAC) plays a role of a gold training set [9] in the area of LOF/GOF retrieval, which well annotates the molecular and cellular events after mutation and captures the semantics of mutation events.

Though the availability of both *mutation type* and *mutation association* offered abundant evidences for gene disease associations, neither single data is empirically rational. GWAS is known for high false positive or negative rate, and the *mutation type* mined from literature only retrieves published knowledge. Therefore, it is illuminative to use text mined *mutation type* to decrease the false positive or negative rate in GWAS, and connect these two heterogeneous mutation data. Since generative model is capable of investigating the distribution of heterogeneous mutation data, it helps to generate and discover the real *de facto* mutations based on both data. In this research, we proposed a generative model, "Gene-Disease Association prediction by Mutation Data Bridging (GDAMDB)", to bridge these two heterogeneous mutation data, and enhance the gene-disease knowledge discovery. The model learns the distribution parameters from *mutation association* data and *mutation type* data, and generating novel data that contains the statistical characteristics of them. Hence, based on the novel data, the model works well to predict novel disease-related genes, which in turn is used for loss/gain-of-function inference and novel gene-disease association discovery.

As an application, GDAMDB was applied to Alzheimer's disease (AD), which is a common neurodegenerative disorder that threaten the elderly for a long time. Although the pathogenesis of AD is still unknown, it is widely accepted that the accumulation of amyloid β forms the plaques on patients' brain thus breaks the calcium equilibrium of the neurons and finally leads the cell apoptosis [10]. Since AD is highly inheritable [11], identifying the important genes sheds light on unveiling the mechanisms of the disease.

The *mutation association* data was downloaded from summary data of the International Genomics of Alzheimer's Project (IGAP) [12], which performed a two-stage GWAS on individuals of European ancestry on 7,055,881 SNPs, and identified 11 new loci of AD. Furthermore, the *mutation types* data was extracted from a PubMed-scale event extraction. After bridging the two heterogeneous mutation data with GDAMDB, we finally retrieved 79 AD associated genes. Intuitively, 2 of them came from the original GWAS study, and 57 of them showed relevance with AD as supported by other GWAS or literature evidences. Therefore, it is believed that GDAMDB successfully combined heterogeneous GWAS and literature data and enhanced the disease-related gene discovery.

2 Methods

In this part, we will introduce the modules and models in the pipeline of GDAMDB model for gene-disease association prediction.

2.1 "*Mutation Type* Retrieval" module in GDAMDB

We designed a "*Mutation Type Retrieval*" module to jointly complete the task of entity recognition and mutation type classification. In this model, we transfer the parameters of BERT-base that released by Google to our model for fine-tuning. Before the input, we use regex and SETH tool for filtering mutation unrelated sentences, the rest of the sentences are thought to be related to *mutation type*. Then, those sentences are used as the input of our joint learning model.

BERT-Base trains model on a large text corpus. For learning deeply bidirectional representation of words, it masks 15% of the words in the input and run the entire sequence through a 12-layer deep bidirectional transformer, each layer is made up of a multi-head self-attention, residual connection, batch normalize and fully connected layer. It updates its parameters though optimizing two tasks, next sentence predicts and mask words predict.

BERT-Base model has been trained by Google, here we transfer its parameters to our joint model. In our model, our input is abstract with sentences filtered, then the sentence is tokenized by WordPiece tool, besides, a classification label ([CLS]) is added in the head of each abstract as classification encoding of each abstract.

After encoding by BERT we get the representation of each word, then a fully connected layer and Softmax are used for normalize classification weights after that CRF loss function is employed to optimize entity recognition task. At the same time, we use word-level attention for the output of the Softmax layer (The output of Softmax layer is regarded as a priori knowledge of mutation type classification). Finally, the attention output vector is concatenate

with the classification label encoded vector, and then a multi-label classifier is used for mutation type classification. Here, we train our model in AGAC train sets, and test out model in development sets. AGAC corpus is designed for extracting gene-mutation type-disease triples from PubMed abstracts.

For filter out SNP related documents in the AD case study, we first use "Alzheimer disease" [MeSH Terms] OR Alzheimer's disease [Text Word]" as the search criteria for downloading 137,473 abstracts from the PubMed database. Then regex and SNP recognize tool SETH was employed to filter out mutation related texts. SETH is an SNP extraction tool for recognizing SNPs and other short sequence variations. Finally, we get 9,430 mutation related abstracts.

The retrieved "mutation types" and part of the sentence evidences are presented in Figure 3. For instance, gene APP is predicted to be related to a GOF *mutation type*, and the sentence evidence is from PubMed with ID equals to 16685645. The sentence, "*...Promoter mutations that increase amyloid precursor-protein expression are associated with Alzheimer disease...*" clearly support the GOF prediction. Moreover, the full results are in the Supplementary file S2 or in online data repository(https://hzaubionlp.com/agac-on-alzheimers-disease/)

2.2 "Synchronization Filter" module in GDAMDB

"Synchronization Filter" module uses a strategy to obtain a gene set with greatest size and most significant literature support. It designs to optimize the probability that most genes in the gene set maintain not only literature significance but also the GWAS significance. Actually, the whole idea of significance integration is an analogue of signal synchronization". Taking this concern, the module is named as "Synchronization Filter" module.

In a mathematical way, we assume there is a gene set for each g with f_{dg}, where f_{dg} is *mutation type* retrieved by "*Mutation Type* Retrieval" module. Since p_{dg} of every g is traceable from GWAS summary data, we order all of the g with f_{dg} with its p-value in a descending order. Generally, the topmost g has greater chance to has *mutation type* info as reported in literature. However, speaking with probability, not all of the g with f_{dg} has greater significance in GWAS. Therefore, from all genes with predicted f_{dg} value, we obtain the top n genes according to their p value, and observe the GWAS significance of this gene set over random set with the same size. The hypothesis test method introduced for this case is Wilcoxon test, where the zero hypothesis H_0 is:

H_0: "The top n genes with literature significance f_{dg} ranked the same with the other genes in GWAS."

Generally, if the p-value obtained by Wilcoxon test is least than a threshold significance value, the H_0 hypothesis will be rejected, and it is accepted that the top n genes with f_{dg} are more significant in GWAS associations if compared with other genes in GWAS. As we hope to get a seed gene set with greater size, the applied strategy is to

increase n gradually from 1 to the maximum. After increment of n, the size of the gene set increases, while p-value of Wilcoxon test specifies the advantage of the gene set over the whole. In most simulation tests, with the increase of n, the plot of $-\log p$ shows a bell shape. This plot suggests that there is a trade-off between the size of gene set and the overall significance over the whole. In the end, a peak value of $-logp$ in Wilcoxon test corresponds a selection of proper value of n, which forms a synchronization filter.

2.3 "SNP-Gene Mapping" module in GDAMDB

For mapping SNP to specific genes. Firstly, we use Bedtools [15] to amplify SNP locations to left and right by 10kb base pairs. Bedtools is a fast, flexible tool set for genome arithmetic, and could be used for SNP flanking creating or calculating overlap of two sets of genomic features. Then, Bedtools is used to map those fragments onto the human genome by chromosomal location. Finally, we use the gene corresponding to the minimum p-value as the mapping result of the SNP.

2.4 Variational Inference on the solution of "Mutation Data Bridging" model in GDAMDB

The following mathematical setup defines the notations and symbols. For a disease d ($d = 1, \cdots, D$) and a gene g ($g = 1, \cdots, G$), $f_{dg} \in \{0,1\}^3$ encodes the associated *mutation type* of gene g for disease d, i.e., LOF/GOF/NA captured from literature, while $p_{dg} \in (0,1)$ refers to the p-value of the mapped *mutation association* of gene g for disease d in GWAS. Both of them are regarded as observations in the graphical model, and marked as dark circle in Figure 1. By introducing a latent variable $\gamma_{dg} \in \{0,1\}$, one switches the significance synchronization of the f_{dg} and p_{dg}. When γ_{dg} switch on, both *mutation type* and *mutation association* show significance; while it switches off, neither does. Thus, significance inconsistency is solved through this synchronization strategy.

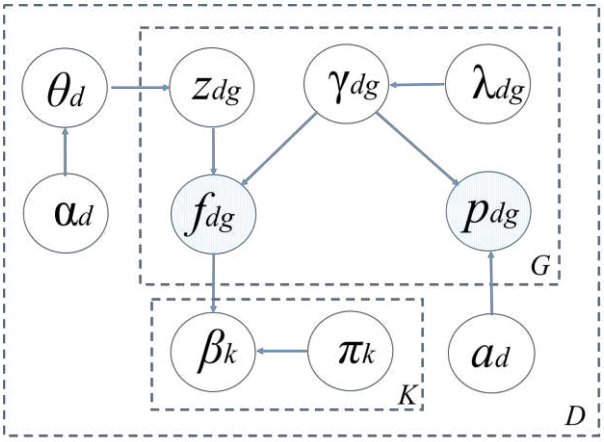

Figure 1: The parameters setting of "Mutation Data Bridging" model.

Referred to Dai et al. [14], a Beta distribution with parameter a_d well cultivate a significant p-value. The greater the a_d, the least the p-value in GWAS. In the meantime, a uniform distribution leads to a non-significance of p-value. Thus, we have

$$p_{dg} \sim \begin{cases} Beta(a_d, 1), & \text{if } \gamma_{dg} = 1; \\ U(0,1), & \text{if } \gamma_{dg} = 0. \end{cases} \tag{3}$$

We assume the switch variable $\gamma_{dg} \sim Bernoulli(\lambda_{dg})$, and $\lambda_{dg} \in (0,1)$. Intuitively, by considering the *mutation type* as event with full semantic augments, a sophisticated language model, Latent Dirichlet Allocation (LDA) [13], is introduced in this model which assumes the *mutation type* f_{dg} is generated by K latent topics, such as interaction, regulation, pathway, molecular or cell physiological activity, etc. Let $z_{dg} \in \{1, \cdots, K\}$ index the latent topic, LDA assumes $z_{dg} \sim Categorical(\theta_d)$, $\theta_d \sim Dir(\alpha_d)$ where $\alpha_d \in \mathbb{R}^K$. Furthermore, the distribution of f_{dg} over the k-th topic is cultivated by a variable $\beta_k \sim Dir(\pi_k)$, $k = 1,2, \cdots, K$, where $\pi_k \in \mathbb{R}^3$. In the meantime, when γ_{dg} equals to zero, a zero vector is assigned to f_{dg}. So we have

$$f_{dg} \sim \begin{cases} Multi(\beta_{z_{dg}}), & \text{if } \gamma_{dg} = 1; \\ \mathbf{0}, & \text{if } \gamma_{dg} = 0. \end{cases} \tag{4}$$

After applying variational inference in this model, the parameters of the distributions for p_{dg} and f_{dg} are computed by the iterative formulas in **Theorem 1** and **Theorem 2**, thus obtain the solutions of the model.

A brief introduction of the solutions is presented in Appendix B, while the complete proofs are shown in supplementary file: File S1. Proof of the generative model of Mutation Data Bridging. The solution of the "Mutation Data Bridging" model is based on **Theorem 1** and **Theorem 2**. For simplicity, all the parameters used in this model are shown as below:

Observation: $F = \{f_{dg}\}, P = \{p_{dg}\}$,

Latent variable: $\theta = \{\theta_d\}, \beta = \{\beta_k\}, \gamma = \{\gamma_{dg}\}, Z = \{z_{dg}\}$,

Model parameter: $\Theta = \{\alpha_d, \pi_k, \lambda_{dg}, a_d\}$.

The first step of the variational inference is to derive an ELBO (Evidence lower bound) by using Jensen's inequality to handle the logarithm of evidence $p(F, P)$.

$$
\begin{aligned}
&\log p\left(\widetilde{F}, P \middle| \Theta\right) \\
&= \log \int_{\theta} \int_{\beta} \sum_{\gamma} \sum_{z} p\left(\widetilde{F}, P, \theta, \beta, \gamma, z \middle| \Theta\right) d\theta d\beta \\
&= \log \int_{\theta} \int_{\beta} \sum_{\gamma} \sum_{z} \frac{p\left(\widetilde{F}, P, \theta, \beta, \gamma, z \middle| \Theta\right)}{q(\theta, \beta, \gamma, z)} \cdot q(\theta, \beta, \gamma, z) d\theta d\beta \\
&\geq E_q\left[\log p\left(\widetilde{F}, P, \theta, \beta, \gamma, z \middle| \Theta\right)\right] - E_q[q(\theta, \beta, \gamma, z)] \\
&:= \text{ELBO}.
\end{aligned} \tag{5}
$$

The variational function 6, is represented by corresponding exponential family distributions.

$$
q(\theta, \beta, \gamma, z) = q(\theta) q(\beta) q(\gamma) q(z) \tag{6}
$$

we assume $q(\theta_d) = Dir(\widetilde{\alpha}_d)$ and $q(\beta_k) = Dir(\widetilde{\pi}_k)$, both of which follow the Dirichlet distribution, while $q(\gamma_{dg}) = Bernoulli(\tilde{\lambda}_{dg})$ follows Bernoulli distribution. In addition, $q(z_{dg}) = Categorical(\widetilde{\theta}_{dg})$ follows a categorical distribution.

Here, $\widetilde{\alpha}$, $\widetilde{\pi}$, $\tilde{\lambda}_{dg}$ and $\widetilde{\theta}_{dg}$ are variational parameters under estimation. For simplicity, we denote $\widetilde{\Theta} = \{\widetilde{\alpha}, \widetilde{\pi}, \tilde{\lambda}_{dg}, \widetilde{\theta}_{dg}\ (d = 1,2, \cdots, D; g = 1,2, \cdots, G)\}$.

Variational parameters are estimated by differential computation of ELBO. And the solution of the model is given by the following theorems.

Theorem 1 Iteration formulas for variational parameters, $\widetilde{\alpha}$, $\widetilde{\pi}$, $\tilde{\lambda}$ and $\widetilde{\theta}$, are:

$$
\begin{aligned}
\widetilde{\alpha}_d^{(t+1)} &= \alpha^{(t)} + \sum_{g=1}^{G} \widetilde{\theta}_{dg}^{(t)} \\
\widetilde{\pi}_k^{(t+1)} &= \pi_k^{(t)} + \sum_{g=1}^{G} \cdot \widetilde{\theta}_{dg;k}^{(t)} \widetilde{f}_{dg}^{(t)} \\
\tilde{\lambda}_{dg}^{(t+1)} &= sigmoid(\log \frac{\lambda_{dg}^{(t)}}{1-\lambda_{dg}^{(t)}} a_d^{(t)} p_{dg}^{(a_d^{(t)}-1)}) \\
\widetilde{\theta}_{dg;k}^{(t+1)} &\propto exp(\sum_{f=1}^{F} \widetilde{f}_{dg;f}^{(t)} (\psi(\widetilde{\pi}_{k;f}^{(t)}) - \psi(\sum_{f=1}^{F} \widetilde{\pi}_{k;f}^{(t)})) \\
&\quad + \psi(\widetilde{\alpha}_{d;k}^{(t)}) - \psi(\sum_{f=1}^{F} \widetilde{\alpha}_{d;k}^{(t)}))
\end{aligned} \tag{7}
$$

Theorem 2 Iteration for computing model parameters, α_d and π_k, is based on Newton's method, $\Theta^{(n+1)} = \Theta^{(n)} - (\mathbf{H}f(\Theta))^{-1} \cdot \nabla f(\Theta)$, where $\mathbf{H}f(\Theta)$ is the Hessian matrix and $\nabla f(\Theta)$ is the gradient of $f(\Theta)$. Then the Newton method iteration of α_d and π_k is based on:

$$
\begin{aligned}
\mathbf{H}L(\alpha_d)_{kj} &= D(\psi'(\textstyle\sum_{k=1}^{K}\alpha_{d;k}) - \delta(k,j)\psi'(\alpha_{d;k})),\\
\nabla L(\alpha_d) &= D(\psi(\textstyle\sum_{k=1}^{K}\alpha_{d;k}) - \psi(\alpha_{d;k}))\\
&\quad + \textstyle\sum_{d=1}^{D}(\psi(\widetilde{\alpha}_{d;k}) - \psi(\sum_{k=1}^{K}\widetilde{\alpha}_{d;k})),\\
\mathbf{H}L(\pi_k)_{fj} &= \psi'(\textstyle\sum_{f=1}^{F}\pi_{k;f}) - \delta(f,j)\psi'(\pi_{k;f}),\\
\nabla L(\pi_k) &= \psi(\textstyle\sum_{f=1}^{F}\pi_{k;f}) - \psi(\pi_{k;f}) + \psi'(\widetilde{\pi}_{k;f})\\
&\quad -\psi'(\textstyle\sum_{f=1}^{F}\widetilde{\pi}_{k;f}).
\end{aligned}
\tag{8}
$$

where $\delta(i,j) = \begin{cases} 1, \text{if } i = j \\ 0, \text{if } i \neq j \end{cases}$

In the meantime, the iteration of model parameters, λ_{dg} and a_d, is

$$
\begin{aligned}
\lambda_{dg}^{(t+1)} &= \tilde{\lambda}_{dg}^{(t)},\\
a_d^{(t+1)} &= -\textstyle\sum_{g=1}^{G}\tilde{\lambda}_{dg}^{(t)} / \sum_{g=1}^{G}(\tilde{\lambda}_{dg}^{(t)}\, logp_{dg})
\end{aligned}
\tag{9}
$$

2.5 Supplementary Files

There are three supplementary files with this research, the name of which are listed as below.

S1. Supplementary file. The whole proof of the mutation data bridging model, Proof of Mutation Data Bridging Model.pdf.

S2. Supplementary data. The result of Mutation type retrieval model: genes with predicted *mutation type* information and the evidence sentence from PubMed texts, Mutation Type Data.xlsx.

S3. Supplementary data. GWAS or literature evidence of 79 predicted AD genes, 79 Predicted Genes.xlsx. The file contains 4 sheets: Sheet 1 is the 12 genes that repeated from IGPA GWAS result, and the three columns are gene id, gene symbol, p-value from IGPA GWAS summary statistics; Sheet 2 is the 24 genes supported by other GWAS, and the four columns are gene id, gene symbol, p-value from the evidence GWAS data, the accession of the evidence GWAS; Sheet 3 are the 33 genes supported by literature, and the 5 columns are gene id, gene symbol, p-value from IGPA GWAS summary statistics, the evidence sentence in literature, the evidence type; Sheet 4 are the newly predicted 10 genes, and the 3 columns are gene id, gene symbol, p-value from IGPA GWAS summary statistics.

3 Results

3.1 Proposed Method "Gene-Disease Association Prediction by Mutation Data Bridging"

3.1.1 Main idea of bridging heterogeneous mutation data

The two kinds of mutation data, *mutation association* and *mutation type*, share the commons but are with different data characteristics, as shown in Figure 2 (a).

First, each mutation data comes from different resources. The *mutation association* is accessible from GWAS summary statistics data, while the *mutation type* is available from literature mining. Second, they both support the investigation of the gene-disease link, but the link is represented by p-value in *mutation association*, and by the confidence value of LOF or GOF from the text mining module in *mutation type*. Third, the evidence of the link is under different concerns. *Mutation association* data are from exacting and reliable experiments, and GWAS are widely accepted as a powerful method to investigate the association between gene and disease. The *mutation type* is only retrieved when a research report the gene are associated with the disease and also describe the mechanism between the gene and disease in a published literature. Forth, both of them also has their weaknesses. For *mutation association*, the high false negative rate and false negative rate in GWAS indicate that not all vital SNPs are able to pass the multiple testing and not all passed genes are real important for the disease. In addition, the lack of consideration about the biological mechanism between genes and phenotype makes GWAS insufficient to pinpoint causal variations. Besides, it is difficult to conduct a GWAS on a large amount of case/control population. For *mutation type*, since it comes from reported literature, it only represents part of the whole knowledge after text mining.

Hence, considering the weaknesses and advantages of these two heterogeneous mutation data, we designed a model to bridge *mutation association* and *mutation type* and achieves data fusion. Generally, a disease casual gene is more likely to be identified by GWAS, and is also more likely to be discovered by other researches and described in the literature. Therefore, bridging *mutation association* and *mutation type* is to integrate the mutation data in a complementing way. In a simplified situation, if the *mutation association* of a significant SNP association failed to pass or barely passed the threshold in GWAS, the *mutation type* of the gene helps to recover the association in the manner of data fusion.

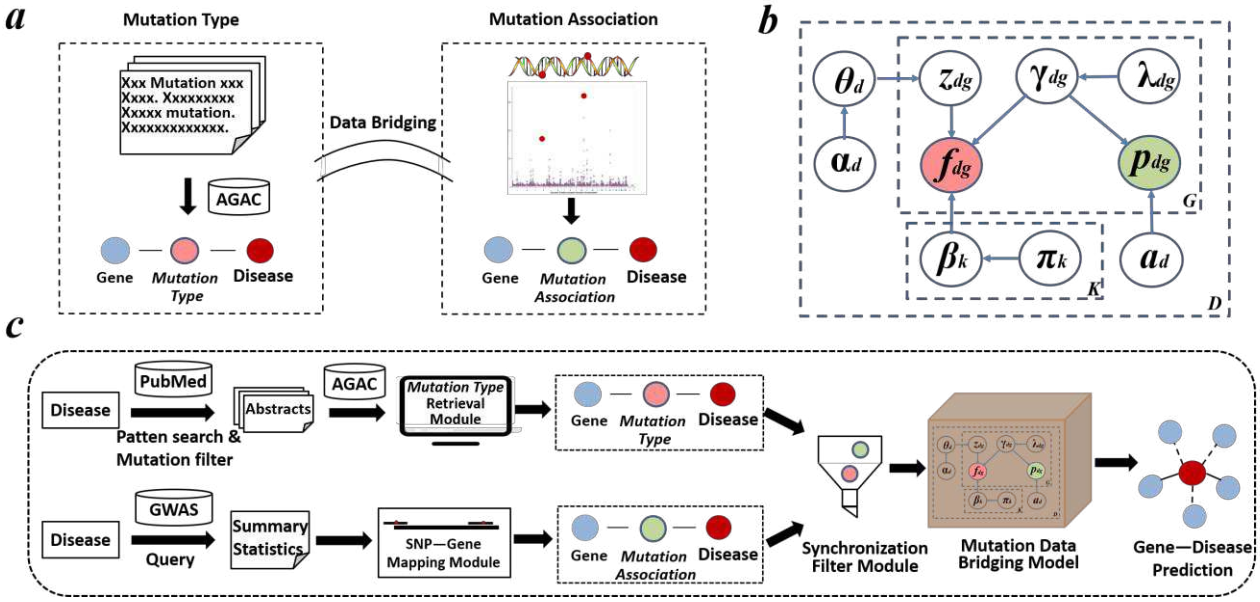

Figure 2: The heterogeneous mutation data and the pipeline of GDAMDB model. (a. The idea of bridging heterogeneous *mutation type* and *mutation association* by using data synchronization. b. Graphical model of the "Mutation Data Bridging" model. c. pipeline of GDAMDB model for gene-disease association prediction.)

3.1.2 Generative model bridges *mutation association* and *mutation type*

We designed a generative model by introducing a switch variable to bridge the *mutation association* and *mutation type* data. Here, the switch variable considers both the significance of *mutation association* mapped to the gene and the reported *mutation type* associated with the gene. Eventually, more reliable disease-related genes are predicted through the integration method.

As shown in Figure 2 (b), the parameter setting of f_{dg} follows the sophisticated language topic model, Latent Dirichlet Allocation (LDA) [13], which automatically organize words to form a readable text learns the probability distribution of each word in different latent topics and so as to generate a new text. By treating gene as the words, our model learns the probability distribution of each gene in different latent topics such as interaction, regulation, pathway, molecular or cell physiological activity, and so forth.

For a given disease d and a gene g, we denote *mutation type* and *mutation association* as f_{dg} and p_{dg} respectively, both of which are regarded as observations in a probability graph from a view of Bayesian statistics.

Referred to Dai et al. [14], a Beta distribution with parameter a_d well cultivate a significant p-value. The greater the a_d, the least the p-value in GWAS. In the meantime, a uniform distribution leads to a non-significance of p-value. Thus we have

$$p_{dg} \sim \begin{cases} Beta(a_d, 1), & \text{if } \gamma_{dg} = 1; \\ U(0,1), & \text{if } \gamma_{dg} = 0. \end{cases} \quad (1)$$

We assume the switch variable $\gamma_{dg} \sim Bernoulli(\lambda_{dg})$, and $\lambda_{dg} \in (0,1)$.

As described above, f_{dg} can be generated by K latent topics in LDA. Let $z_{dg} \in \{1, \cdots, K\}$ index the latent topic, LDA assumes $z_{dg} \sim Categorical(\theta_d)$, $\theta_d \sim Dir(\alpha_d)$ where $\alpha_d \in \mathbb{R}^K$. Furthermore, the distribution of f_{dg} over the k-th topic is cultivated by a variable $\beta_k \sim Dir(\pi_k)$, $k = 1,2, \cdots, K$, where $\pi_k \in \mathbb{R}^3$. In the meantime, when γ_{dg} equals to zero, a zero vector is assigned to f_{dg}. So we have

$$f_{dg} \sim \begin{cases} Multi(\beta_{z_{dg}}), & \text{if } \gamma_{dg} = 1; \\ \mathbf{0}, & \text{if } \gamma_{dg} = 0. \end{cases} \quad (2)$$

When a gene g mutated and the mutation related to disease d, γ_{dg} equals to 1. Decided by the value of γ_{dg}, the $p_{dg} \sim Beta(a_d, 1)$ which will mostly be a small value and the $f_{dg} \sim Multi(\beta_{z_{dg}})$ which will mostly be 1. The values are consisting with the facts that the gene should be significant in the GWAS result of this disease and the description about the gene mutation can be found in the literature of this disease.

After applying variational inference in this model, the parameters of the distributions for p_{dg} and f_{dg} are computed by the iterative formulas, thus obtain the solutions of the model. Definition of all the distribution parameters in the figure is provided in METHODS (Section 2.4), while a complete proof of the generative model and deduction can be found in the supplementary file, S1. Proof of Mutation Data Bridging Model.pdf.

3.1.3 Pipeline of "Gene-Disease Association prediction by Mutation Data Bridging"

The pipeline of "Gene-Disease Association prediction by Mutation Data Bridging (GDAMDB)" is shown in Figure 2(c), which consists of three data processing modules and one prediction model, i.e., "*Mutation Type* Retrieval" module, "SNP-Gene Mapping" module, "Synchronization Filter" module, and "Mutation Data Bridging" model.

(1) "*Mutation Type* Retrieval" is a text mining module, which is capable of jointly recognizing the gene and other entities from literature and classifying the *mutation type* of the genes based on the semantic in the sentence. The module is designed based on a Bidirectional Encoder Representations from Transformers (BERT) model which is released by Google and trained on a large text corpus. BERT contains 12 layers of deep bidirectional transformer, and each of them is fully connected and made up of a multi-head self-attention, residual connection, batch normalize. The complex construction makes BERT able to lean deep bidirectional representation of each

word. Therefore, we transfer and fine-tuned its parameters on our joint model. As shown in Figure 2 (c), after paternal searching from PubMed and mutation filtering, the abstracts containing diseases and mutations are input into BERT, then the presentations of each word in the abstracts are obtained. Subsequently, a fully connected layer and softmax are used to normalize classification weights, and CRF loss function is employed to optimize entity recognition task in the meantime. Finally, the model output the *mutation type* of genes in abstracts.

(2) "SNP-Gene Mapping" module is to process the *mutation association* data. Since the pipeline is focus on gene, the SNPs in GWAS data should be mapped on genes by bedtools [15]. The p-value of a gene is the p-value of its SNP which is the lowest one.

(3) "Synchronization Filter" module designs to optimize the probability that most genes in the gene set maintain not only literature significance but also the GWAS significance. Generally, the gene with great significant *mutation association*, p_{dg}, in GWAS is likely to be described in literature with *mutation type*, f_{dg}, but not all of genes satisfy the rule. Therefore, from all genes with predicted f_{dg} value, we obtain the top n genes according to their p value, and observe the GWAS significance of this gene set over random set with the same size. The hypothesis test method introduced for this case is Wilcoxon test.

(4) "Mutation Data Bridging" model is the generative model mentioned above which bridges *mutation association* data and *mutation type* data by introducing a switch variable. The inputs of this model are the two processed mutation data and the gene set selected by "Synchronization Filter" module. Then the distributions of the two mutation data are computed by the model, and the predicted gene are obtained based on the switch variable of each gene.

Briefly speaking, f_{dg} is retrieved from PubMed by using the "*Mutation Type* Retrieval" model and AGAC corpus, and genes with significant f_{dg} are defined as "literature significant" genes. In the meantime, *mutation association*, p_{dg}, is extracted from GWAS summary data by applying SNP inclusion criteria, and "GWAS significant" genes are obtained by using a "SNP-Gene Mapping" module. In order to better synchronize the above heterogeneous mutation data, a "Synchronization Filter" module creates a seed gene set consists of g with significant $\widehat{f}_{dg}$ and p_{dg}. After feeding the observations, $\{\widehat{f}_{dg}\}$ and $\{p_{dg}\}$, into the "Mutation Data Bridging" model, the model parameters are obtained. Eventually, generative process is carried on to produce novel *mutation types* with significant $\widetilde{f}_{dg}$. Thus,

new appeared gene g with $\tilde{f}_{dg}$ is predicted with novel gene-disease association. All of the pipeline details are elucidated in Online Method.

The purpose of GDAMDB is to accelerate the discovery rate of the gene associations of GWAS by integrating both *mutation association* and *mutation type* information. A case study on Alzheimer's disease (AD) was carried on to evaluate the performance of GDAMDB in the support of discovery of novel gene-disease associations.

3.2 Application of GDAMDB on Alzheimer's Disease

Alzheimer's Disease is a common neurodegenerative disorder, which impairs the memory, language and various body behaviors. Till now, there are 116 AD studies in GWAS Catalog [16], three of which provide the summary statistics data. We chose one of the studies that provided the most complete data. Although no database recording the *mutation type* info of AD-related genes, there are lots of literatures that report the studies of AD pathogenesis. The *mutation type* of the genes is widely implied in the description of the literatures. Since AD is a important diseases and the mutation data of AD are available, we apply GDAMDB on AD to retrieve the genes that are undiscovered.

3.2.1 *Mutation association* data of Alzheimer's disease

In 2013, the International Genomics of Alzheimer's Project (IGAP) [12] performed a two-stage GWAS on individuals of European ancestry on 7,055,881 SNPs. In stage 1, they meta-analyzed four previous AD GWAS datasets including 17,008 AD cases and 37,154 controls. In stage 2, they tested 211,632 SNPs on 8,572 AD cases and 11,312 controls. The final result was obtained after meta-analysis of combining stage 1 and 2. We selected the summary statistics file that combined stage 1 and 2, in which contains 1,513 genes. The p-value of each gene was same with the most significant SNP in the gene.

3.2.2 *Mutation type* data of Alzheimer's disease

In the meantime, the MeSH term "Alzheimer's disease" was used as the key word to query PubMed database, and 137,473 abstracts were downloaded. To ensure that the literatures contain description about mutation, SETH [17] was applied to filter the literature. SETH is able to recognize the SNP or other mutation semantic words in texts. Thus, till this procedure, the abstracts that contains AD and mutations were left. After mutation filtering, 9,430 abstracts were input into "*Mutation Type* Retrieval" module. The module will compute the confidence value for each abstract in each *mutation type*. The output of the module is the *mutation type* of genes in each abstract, of which the confidence value passes the module threshold. Subsequently, we manually checked the result, and only preserved the abstracts that clearly describe the mechanism of a mutated gene leading to AD. Finally, 65 genes with their

mutation types are firstly recognized from 325 abstracts where each abstract contains a conclusive sentence evidence leading to a *mutation type*. It is noted that the obtained plenty of AD-related LOF/GOF data is new to the AD community, while the full result is offered in supplementary data, S2: genes with predicted *mutation type* information and the evidence sentence from PubMed texts, Mutation Type Data.xlsx.

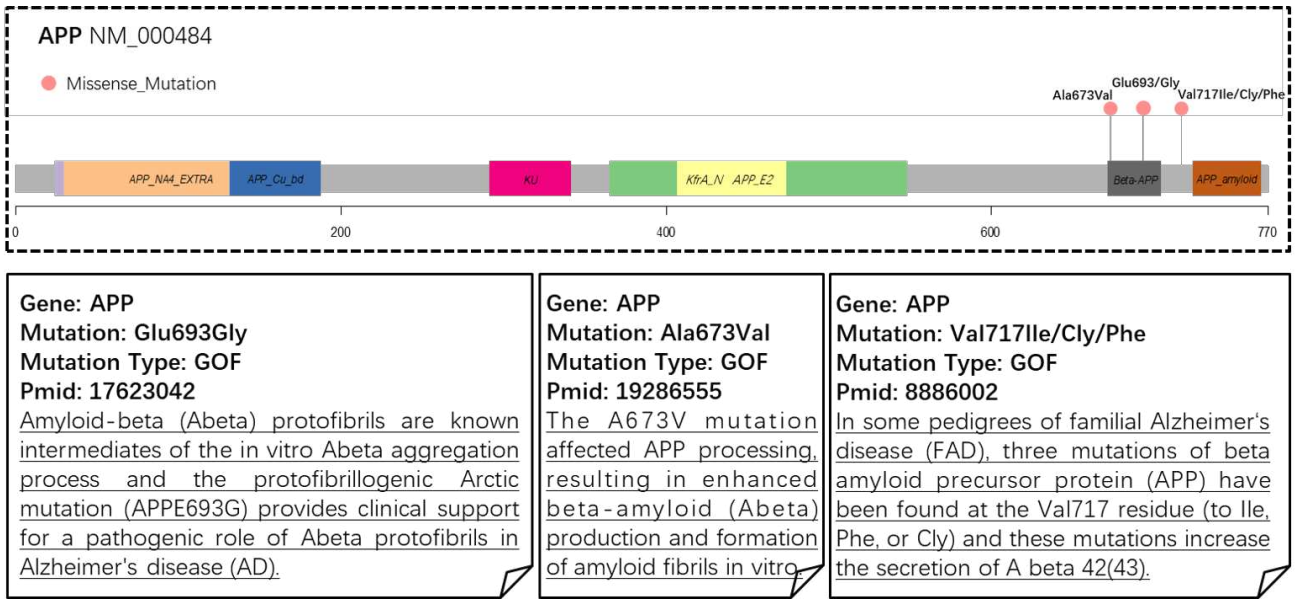

Figure 3: The illustrated examples of *mutation types* retrieved from Glu693Gly, Ala673Val, and Val717Ile/Cly/Phe in APP.

56 abstracts clearly described the amino acid change of the mutations or the rs number of the SNPs, which totally report 64 mutations on 28 genes. Among the 64 mutations, 54 of them locate on the coding regions of the genes leading to the amino changes, and 10 of them locate on the non-coding region of the genes. Figure 3 is an example, which shows the specific mutations of APP recognized by the module. These five mutations are on three locations of APP, two of which locate on the Beta-APP domain of APP protein. The amino acid location 717 is found three mutations, and this location is between the sequence producing Beta-APP domain and sequence producing APP amyloid domain which form the beta-amyloid and is strongly implicated in the pathogenesis of AD. Moreover, the corresponding sentence evidences of these APP mutations are below. As introduced above, this module is able to recognize the entities and classify the *mutation type* of a gene. For example, the sentence in the middle, "*The A673V mutation affected APP processing, resulting in enhanced beta amyloid (Abeta) production and formation of amyloid fibrils in vitro.*", *APP* and *enhanced* will be recognized by the module. Based on *enhanced*, the confidence value of GOF will be higher than the value of LOF in this abstract, hence the *mutation type* of APP will be classified as GOF. Therefore, among the 325 abstracts, each one are recognized at least one gene and their *mutation types*.Besides that, all the 325 abstracts carry the clear semantic of the downstream biological processes after mutation, which can be

divided into 8 types after manual curation. As shown in figure 4, Gene Expression, Protein Activity, Interaction, Pathway Activity and Cell Activity are the fundamental biological processes which follows the central dogma and are from molecular level to cell level. In addition, the Phosphorylation, Abeta Accumulation and Ca2+ Concentration are frequently mentioned. Interestingly, these three biological processes are related to the known hypothesizes of AD pathogenesis. Abeta is the production of APP gene, the accumulation of which, especially Abeta42, forms the fibrillar amyloid plaques in brain and impair the ability of spatial learning and memory [18]. Phosphorylation related to another hypothesis of AD pathogenesis, especially the phosphorylation of Tau protein which encoded by MAPT gene. The hyperphosphorylation of Tau protein leads to neurofibrillary tangles in neurons and eventually results in the apoptosis of neurons [19]. Intracellular Ca2+ concentration is also thought as part of the cause of AD. The dysregulation of intracellular Ca2+ signaling disturbs many neural processes, which implicated in AD mechanism [20].

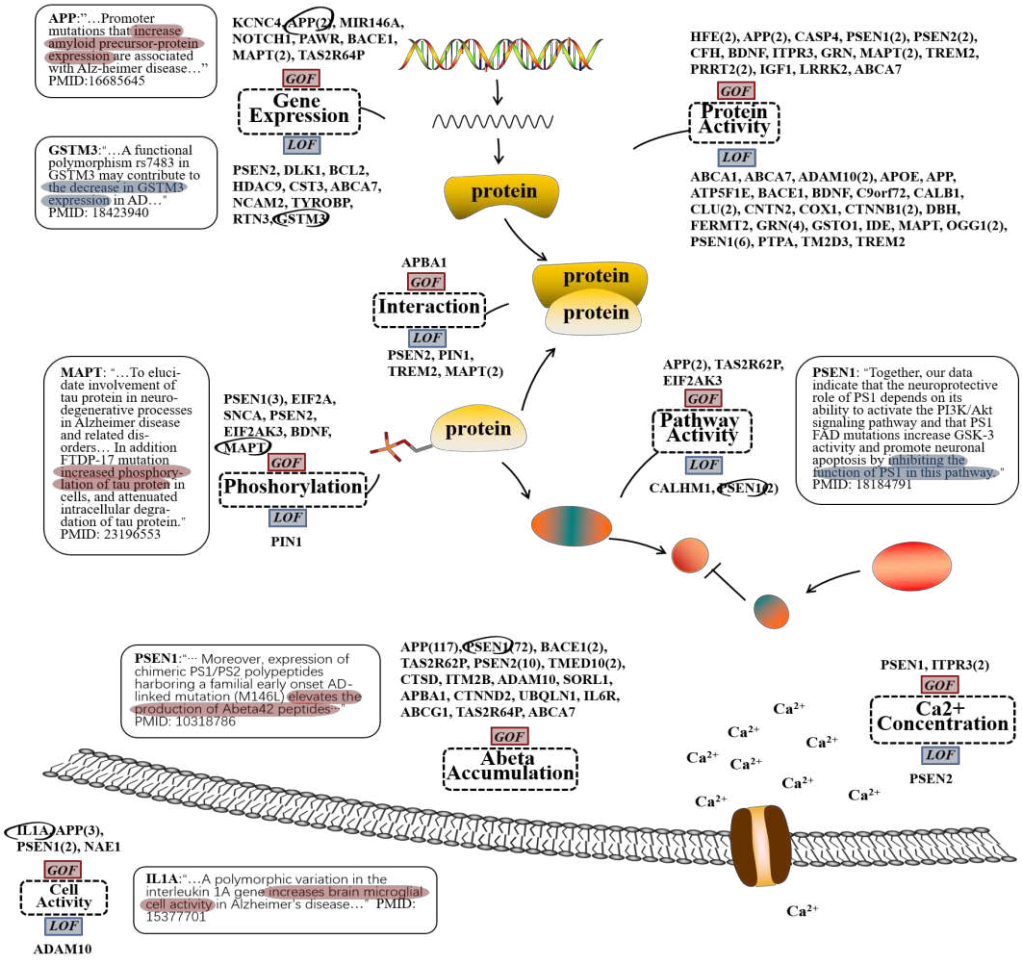

Figure 4: Biological process categories of 325 mutation types and the sentence evidences.

As shown in the upper left of the figure 4, it presents the genes that related to the gene expression biological process. the mutations in the 8 genes, KCNC4, APP, MIR146A, NOTCH1, PAWR, BACE1, MAPT, TAS2R64P, are GOF

mutation and leads to the increase of gene expression. In addition, the increase of expression pf APP and MAPT are mentioned in two abstracts. 10 genes are mentioned LOF mutation that decrease the gene expression in abstracts. At the left side of the genes are two sentence examples. In the sentence, "*...Promoter mutations that increase amyloid precursor-protein expression are associated with Alzheimer disease...*", "*increase*" helps to confirm GOF and "*amyloid precursor protein expression*" helps to confirm that the biological process that effected by mutation is gene expression. Similarly, GSTM3 is grouped into the LOF of gene expression.

The biological process category of the genes and evidence sentence can be found in supplementary data S2, or in online data repository (https://hzaubionlp.com/agac-on-alzheimers-disease/)

3.2.3 Data fusion of heterogeneous mutation data

The data fusion by GDAMDB is shown on Figure 4. The left two graphs present confidence value of the gene *mutation type* in each abstract and the p-value of gene *mutation association*, both of the graphs are showing the rough distributions of data. In *mutation type* graph, the 325 gene-*mutation types*-AD are predicted and manually checked from 9,430 abstracts, and there are 65 unique genes since some of the genes are mentioned more than once in these abstracts. The empirical threshold represents the model parameters and human filtering. In *mutation association* graph, there are 23 *mutation associations* passed the final Bonferroni threshold, and they are mapped to 23 genes. The graph became bar graph after data fusion, since the output of the model is binary info representing the association between gene and disease or not. 79 genes are predicted to be the AD-related genes. The final prediction filtered some of the genes that passed the threshold in the single mutation data but recognized as the false negative genes by the model, and also retrieved the genes failed to pass the threshold in the single mutation data but recognized as the AD-related genes.

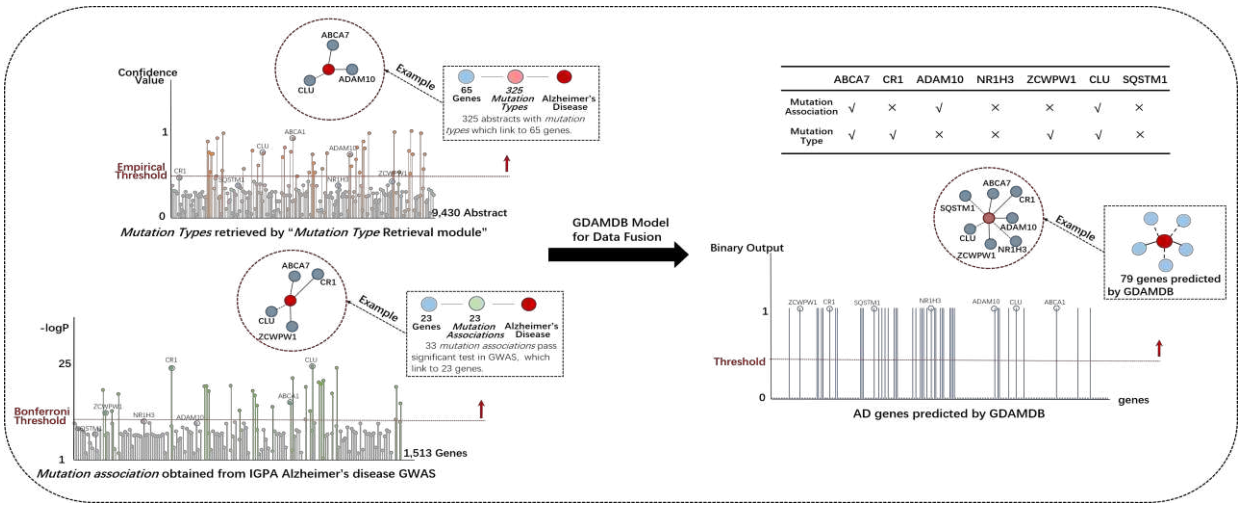

Figure 5: Data fusion of heterogeneous AD mutation data improves the discovery of novel AD-related genes.

For example, as shown in the circle above *mutation types* graph, the *mutation types* of the three genes are retrieved by "mutation *type* retrieval module" and passed the empirical threshold, ABCA7, CLU and ADAM10. The circle above the *mutation association* graph contains four genes that passed the Bonferroni threshold, ABCA7, CLU, CR1 and ZCWPW1. There are different limitations make the information that mutation data contained is incomplete. Therefore, as marked on the graphs, ABCA7 and CLU both pass the threshold in two kinds of mutation data, but ADAM10, ZCWPW1 and CR1 only pass one. However, the -logp value of ADAM10 is close to the Bonferroni threshold, while the confidence value of ZCWPW1 and CR1 are close to the empirical threshold.

After data fusion, ABCA7, CLU, ADAM10, CR1 and ZCWPW1 are output by GDAMDB, which shows that GDAMDB is able to break the limitation of these two mutation data and save the important genes that are failed to pass the threshold. Besides, the genes, NR1H3 and SQSTM1, are retrieved in neither *mutation type* data nor *mutation association* data, but retrieved by GDAMDB after data fusion. It shows that GDAMDB is not simply merge the genes that are significant in one of the mutation data, but to learn the latent regularity of the mutation data distribution.

3.2.4 Novel discovery of AD-related genes after heterogeneous mutation data fusion

An encouraging result of AD-related gene discovery is shown in Figure 6. The left ellipse refers to the significant gene set which is reported in IGAP GWAS research [12]. Meanwhile, the right ellipse represents 79 genes that are predicted by GDAMDB.

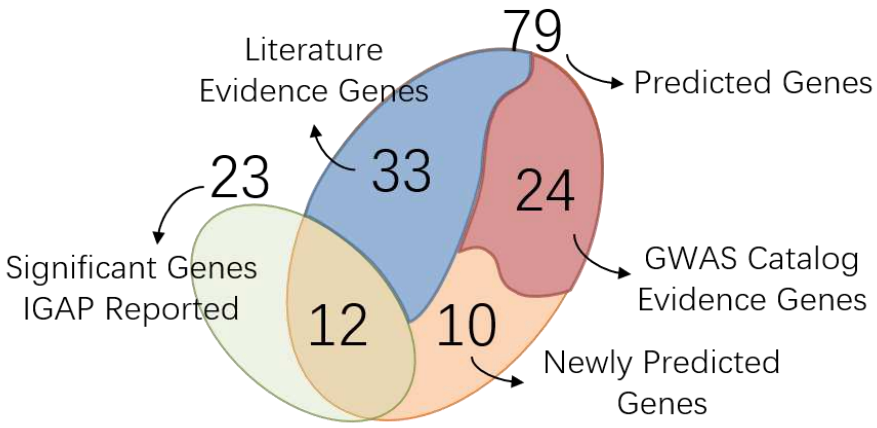

Figure 6: 69 out of 79 predicted genes have supportive AD-related evidence.

As shown in Figure 6, 12 out of 79 predicted genes are reported in IGAP GWAS research, which are CR1, CD2AP, EPHA1, CLU, PICALM, ABCA7, HLA-DBR1, PTK2B, SORL1, INPP5D, ZCWPW1 and FERMT2. The red part of the big ellipse contains 24 genes with GWAS catalog evidences, which are reported to be AD-related in GWAS catalog

(https://www.ebi.ac.uk/gwas/efotraits/EFO_0000249). They are RNU6560P, GULOP, EPHA1-AS1, STAG3L5P-PVRIG2P-PILRB, HLA-DQA1, GPR141, ADGRF2, SCARA3, SPI1, CSTF1, AP4M1, SCIMP, PILRA, EPDR1, RAPSN, MS4A6E, PSMC3, CASTOR3 and MS4A2 from GCST5922, HLA-DRB9, ADAM10 from GCST007320, MEF2C-AS1 from GCST003427, NDUFAF6 from GCST009021, and MADD from GCST007825.

As for the other 43 genes which are not recorded in GWAS catalog as AD-related genes, 33 of them are reported to be AD-related genes with confirmed literature evidence. In details, NYAP1, MYBPC3, BTNL2, HLA-DQB1, AGBL2, COPS6, ZKSCAN1, MTCH2, MCM7, NDUFS3, PILRB, SLC39A13, USP50 and ACP2 [21] are identified as the additional genes within the loci which contain AD-related SNP, while ADGRF4 [22] and RNU6-603P [23] are the gene adjacent to AD-related SNP. MS4A4E [24], SLC25A1P1 [25], MBLAC1 [26], USP8 [27], TSPOAP1 [28], PGF [29], NR1H3 (also named as LXRα) [30] , TP53INP1 [31] and TRIP4 [32] are reported in AD GWAS study or the research that identified new genes by performing further experiment on AD GWAS data. There are 5 microRNAs in the prediction gene set. MicroRNA is known to regulate neuronal development, and the abnormal alteration of which contributes to neurodegenerative disorders. The 5 microRNA are also found to be related to AD or neurodegenerative disease, which are MIR-4487 [33], MIR142 [34], MIR25[35], MIR4736 [36] and MIR106B [37]. Besides, MAML1 [38], CR2 [39] and SQSTM1 [40] are reported to be related to neurodegenerative disease.

The rest of the 10 genes are all pseudogenes, which are IGHVIII-67-4, MTCO3P1, RNA5SP340, RN7SKP116, SNORD3P2, SNORD3P3, SNORD3P1, RNU5E-10P, IGHV3-71 and IGHV2-70. Although there is no direct GWAS Catalog evidence or literature evidence showing that these genes are associated with AD, the regulation function of the pseudogenes in the pathogenesis of complex disease is worth attention. In 2011, Tay et al. [41] firstly proposed the hypothesis that the competing endogenous RNAs (ceRNAs) may compete microRNA with the mRNA that contains same microRNA response element with the ceRNAs. Costa et al. [42] suggested that ceRNAs may be involved in the pathogenesis of neurodegenerative disease, such as AD.

Among the 11 genes that are reported in IGAP but aren't retrieved by GDAMDB, most of them are less reported by GWAS Catalog. Especially, 4 of them are only reported once by IGAP in whole GWAS Catalog.

In summary, among the 79 predicted genes, the genes reported by IGAP, the genes with AD GWAS catalog evidences, and the genes with AD literature evidences accounts for 86% (69/79), and the rest 14% (10/79) predicted genes are suggested to be potentially involved in the parthenogenesis mechanism of AD.

4 Discussion

In the case of data fusion in terms of knowledge discovery, the knowledge can be any form of data with different format. In our research, the association relation between gene and disease can be the p-value, named as *mutation association*, in GWAS, where the smaller p-value represents the more significant relevance between gene and disease. Furthermore, this association relation can also be *mutation type* in literature, where the description about the mechanism of mutations in disease pathogenesis directly indicate the details of the relation. When different data reveal the relations in different aspects, taking both aspects into consideration leads to a more comprehensive knowledge discovery. Besides, since the advantages and weakness vary from heterogeneous data, data fusion helps to enhance the quality of both data. The relevance between a gene and a disease is adjusted after data fusion, especially when a false negative *mutation association* of a gene fails to pass significant test in GWAS but is found to be active with *mutation type* information in literature.

A generative model is capable of learning the data distribution of observations from two heterogeneous categories, and generating novel data which represents the statistical characteristics of both observations, thus achieves the data fusion of heterogeneous data. Therefore, by bridging *mutation type* data and *mutation association* data, GDAMDB is capable of retrieving the important AD-related genes that are failed to pass the multiple testing in GWAS or haven not been reported in literature. Eventually, our model retrieved 79 AD genes, and 57 of them are not reported in the source GWAS study but 47 out of 57 are supported by convince evidences that are AD-related genes, which positively shows the reliability of the model performance.

As a generative model, GDAMDB offers a way to enhance the disease-related gene discovery in a single mutation data, and the implementation procedure of the model shows that the model is flexible to be adopted to each given disease, in the case when the GWAS summary data and sufficiently abundant literature are available. All the results in this research indicate that data fusion sheds light to the novel knowledge discovery.

5 Conclusion

This research drew a novel respective towards the data form of mutations, of which there are mutation associations obtained from GWAS experiment and mutation type extracted from text mining. It is known that GWAS associations are under-powered for pinpointing causal genes due to high false positive/negative rate, and integration of other mutation information is possibly an effective addition. Thus, we used a PubMed-wide text mining strategy to pinpoint vital genes which carry core semantics of the mutation effect, and came up with the mining of *mutation type*, which associated gene and disease in an interpretable LOF/GOF way.

GDAMDB is a model to bridge the two heterogeneous types of mutation data from a same disease. This model designs a switch variable γ_{dg} to synchronize the importance of two types of mutation data, learns the distribution of both data, and discover novel significant genes that potentially related to disease.

The case study in AD made use of real GWAS data from IGAP and went through a thorough data integration via applying GDAMDB. Finally, 57 out of 79 predicted genes are closely related to AD. The results obtained in this research fully showed that bridging the heterogeneous mutation data integrate information from GWAS and literature, thus shed lights on novel disease-related gene discovery.

6 Abbreviations

Abbreviations used in this paper are listed as below.

AD, Alzheimer's disease
AGAC, active gene annotation corpus
APP, amyloid precursor protein
BERT, bidirectional encoder representations from transformers
GDAMDB, gene-disease association prediction by mutation data bridging
GOF, gain of function
GWAS, genome-wide association study
IGAP, international genomics of Alzheimer's project
LDA, latent Dirichlet allocation
LOF, loss of function
SNP, single nucleotide polymorphism

7 Declaration

7.1 Ethics approval and consent to participate

This research does not involve human or animals.

7.2 Consent for publication

All the authors have consented for the publication.

7.3 Availability of data and material

Data and materials are available in the supplementary files.

7.4 Competing interests

None of the authors have any competing interests.

7.5 Funding

The research is supported by Hubei Province Funds for Natural Science (No. 2019CFB552).

7.6 Authors' contributions

KZ developed the algorithm and coding. YW checked the biological knowledge and wrote the manuscript. KBC and JDK took part in the algorithm discussion and the manuscript writing. XM took part in the algorithm development. ZS curated the biological data. XM checked the biological data. JX designed the whole pipeline and took responsibility of the whole research. All authors read and approved the final manuscript.

7.7 Acknowledgements

The authors would like to express their gratitude to all HZAU BioNLP teams members who joined many discussions related to this research.

References

[1] P. Jia, S. Zheng, J. Long, W. Zheng, and Z. Zhao. dmgwas: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, 27(1):95–102.

[2] Andrea Califano, Atul J Butte, Stephen Friend, Trey Ideker, and Eric Schadt. Leveraging models of cell regulation and gwas data in integrative network-based association studies. *Nature genetics*, 44(8):841, 2012.

[3] Kai Wang, Mingyao Li, and Hakon Hakonarson. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854.

[4] Yongsheng Li, Yunpeng Zhang, Xia Li, Song Yi, and Juan Xu. Gain-of-function mutations: an emerging advantage for cancer biology. *Trends in biochemical sciences*, 2019.

[5] Zhong Yi Wang and Hong Yu Zhang. Rational drug repositioning by medical genetics. *Nature biotechnology*, 31(12):1080, 2013.

[6] Martin Citron, David Westaway, Weiming Xia, George Carlson, Thekla Diehl, Georges Levesque, Kelly Johnson-Wood, Michael Lee, Peter Seubert, Angela Davis, et al. Mutant presenilins of alzheimer's disease increase production of 42-residue amyloid β-protein in both transfected cells and transgenic mice. *Nature medicine*, 3(1):67–72, 1997.

[7] Rita Guerreiro, Aleksandra Wojtas, Jose Bras, Minerva Carrasquillo, Ekaterina Rogaeva, Elisa Majounie, Carlos Cruchaga, Celeste Sassi, John SK Kauwe, Steven Younkin, et al. Trem2 variants in alzheimer's disease. *New England Journal of Medicine*, 368(2):117–127, 2013.

[8] Stacy Steinberg, Hreinn Stefansson, Thorlakur Jonsson, Hrefna Johannsdottir, Andres Ingason, Hannes Helgason, Patrick Sulem, Olafur Th Magnusson, Sigurjon A Gudjonsson, Unnur Unnsteinsdottir, et al. Loss-of-function variants in abca7 confer risk of alzheimer's disease. *Nature genetics*, 47(5):445–447, 2015.

[9] Yuxing Wang, Kaiyin Zhou, Jin Dong Kim, Kevin Bretonnel Cohen, Mina Gachloo, Yuxin Ren, Shanghui Nie, Xuan Qin, Panzhong Lu, and Jingbo Xia. An active gene annotation corpus and its application on anti-epilepsy drug discovery. *BIBM 2019: International Conference on Bioinformatics & Biomedicine, San Diego, U.S*, 2019.

[10] Dennis J Selkoe. Translating cell biology into therapeutic advances in alzheimer's disease. *Nature*, 399(6738):A23–A31, 1999.

[11] Robert S Wilson, Sandra Barral, Joseph H Lee, Sue E Leurgans, Tatiana M Foroud, Robert A Sweet, Neill Graff-Radford, Thomas D Bird, Richard Mayeux, David A Bennett, et al. Heritability of different forms of memory in the late onset alzheimer's disease family study. *Journal of Alzheimer's Disease*, 23(2):249–255, 2011.

[12] Jean Charles Lambert, Carla A Ibrahim Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, C´eline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 45(12):1452, 2013.

[13] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[14] Mingwei Dai, Jingsi Ming, Mingxuan Cai, Jin Liu, Can Yang, Xiang Wan, and Zongben Xu. Igess: a statistical approach to integrating individual-level genotype data and summary statistics in genome-wide association studies. *Bioinformatics*, 33(18):2882–2889, 2017.

[15] Aaron R Quinlan. Bedtools: the swiss-army tool for genome feature analysis. *Current protocols in bioinformatics*, 47(1):11–12, 2014.

[16] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

[17] Philippe Thomas, Tim Rockt¨aschel, J¨org Hakenberg, Yvonne Lichtblau, and Ulf Leser. Seth detects and normalizes genetic variants in text. *Bioinformatics*, 32(18):2883–2885, 2016.

[18] John Hardy and David Allsop. Amyloid deposition as the central event in the aetiology of alzheimer's disease. *Trends in pharmacological sciences*, 12:383–388, 1991.

[19] Wanjoo Chun and GV Johnson. The role of tau phosphorylation and cleavage in neuronal cell death. *Front Biosci*, 12(733):e56, 2007.

[20] Frank M LaFerla. Calcium dyshomeostasis and intracellular signalling in alzheimer's disease. *Nature Reviews Neuroscience*, 3(11):862–872, 2002.

[21] Anastasia G Efthymiou and Alison M Goate. Late onset alzheimer's disease genetics implicates microglial pathways in disease risk. *Molecular neurodegeneration*, 12(1):43, 2017.

[22] Anna A Pimenova, Towfique Raj, and Alison M Goate. Untangling genetic risk for alzheimer's disease. *Biological psychiatry*, 83(4):300–310, 2018.

[23] Brian W Kunkle, Benjamin Grenier-Boley, Rebecca Sims, Joshua C Bis, Adam C Naj, Anne Boland, Maria Vronskaya, Sven J van der Lee, Alex Amlie-Wolf, Celine Bellenguez, et al. Meta-analysis of genetic association with diagnosed alzheimer's disease identifies novel risk loci and implicates abeta, tau, immunity and lipid processing. *bioRxiv*, page 294629, 2018.

[24] Paul Hollingworth, Denise Harold, Rebecca Sims, Amy Gerrish, Jean Charles Lambert, Minerva M Carrasquillo, Richard Abraham, Marian L Hamshere, Jaspreet Singh Pahwa, Valentina Moskvina, et al. Common variants at abca7, ms4a6a/ms4a4e, epha1, cd33 and cd2ap are associated with alzheimer's disease. *Nature genetics*, 43(5):429, 2011.

[25] Shubhabrata Mukherjee, Joshua C Russell, Daniel T Carr, Jeremy D Burgess, Mariet Allen, Daniel J Serie, Kevin L Boehme, John SK Kauwe, Adam C Naj, David W Fardo, et al. Systems biology approach to lateonset alzheimer's disease genome-wide association study identifies novel candidate genes validated using brain expression data and caenorhabditis elegans experiments. *Alzheimer's & Dementia*, 13(10):1133–1142, 2017.

[26] Iris J Broce, Chin Hong Tan, Chun Chieh Fan, Iris Jansen, Jeanne E Savage, Aree Witoelar, Natalie Wen, Christopher P Hess, William P Dillon, Christine M Glastonbury, et al. Dissecting the genetic relationship between cardiovascular risk factors and alzheimer's disease. *Acta neuropathologica*, 137(2):209–226, 2019.

[27] Eniola Funmilayo Aduke Yeates and Giuseppina Tesco. The endosome-associated deubiquitinating enzyme usp8 regulates bace1 enzyme ubiquitination and degradation. *Journal of Biological Chemistry*, 291(30):15753–15766, 2016.

[28] Gyungah R Jun, Jaeyoon Chung, Jesse Mez, Robert Barber, Gary W Beecham, David A Bennett, Joseph D Buxbaum, Goldie S Byrd, Minerva M Carrasquillo, Paul K Crane, et al. Transethnic genome-wide scan identifies novel alzheimer's disease loci. *Alzheimer's & Dementia*, 13(7):727–738, 2017.

[29] Pu-Ting Xu, YiJu Li, Xue-Jun Qin, Clemens R Scherzer, Hong Xu, Donald E Schmechel, Christine M Hulette, John Ervin, Steven R Gullans, Jonathan Haines, et al. Differences in apolipoprotein e3/3 and e4/4 allele-specific gene expression in hippocampus in alzheimer disease. *Neurobiology of disease*, 21(2):256–275, 2006.

[30] Noam Zelcer, Negar Khanlou, Ryan Clare, Qingguang Jiang, Erin G Reed-Geaghan, Gary E Landreth, Harry V Vinters, and Peter Tontonoz. Attenuation of neuroinflammation and alzheimer's disease pathology by liver x receptors. *Proceedings of the National Academy of Sciences*, 104(25):10601–10606, 2007.

[31] Valentina Escott-Price, C´eline Bellenguez, Li-San Wang, Seung-Hoan Choi, Denise Harold, Lesley Jones, Peter Holmans, Amy Gerrish, Alexey Vedernikov, Alexander Richards, et al. Gene-wide analysis detects two new susceptibility genes for alzheimer's disease. *PloS one*, 9(6):e94661, 2014.

[32] A Ruiz, S Heilmann, Tim Becker, Isabel Hern´andez, Hermann Wagner, M Thelen, A Mauleon, M RosendeRoca, C´eline Bellenguez, JC Bis, et al. Follow-up of loci from the international genomics of alzheimer's disease project identifies trip4 as a novel susceptibility gene. *Translational psychiatry*, 4(2):e358, 2014.

[33] Ling Hu, Rong Zhang, Qiong Yuan, Yinping Gao, Mary Q Yang, Chunxiang Zhang, Jiankun Huang, Yufei Sun, William Yang, Jack Y Yang, et al. The emerging role of microrna-4487/6845-3p in alzheimer's disease pathologies is induced by aβ25–35 triggered in sh-sy5y cell. *BMC systems biology*, 12(7):119, 2018.

[34] Juhyun Song and Young-Kook Kim. Identification of the role of mir-142-5p in alzheimer's disease by comparative bioinformatics and cellular analysis. *Frontiers in molecular neuroscience*, 10:227, 2017.

[35] Linlin Wang, Li Min, Qingdong Guo, Junxia Zhang, Hailun Jiang, Shuai Shao, Jianguo Xing, Linlin Yin, Jianghong Liu, Rui Liu, et al. Profiling microrna from brain by microarray in a transgenic mouse model of alzheimer's disease. *BioMed research international*, 2017, 2017.

[36] Masataka Katsu, Yuka Hama, Jun Utsumi, Ken Takashina, Hiroshi Yasumatsu, Fumiaki Mori, Koichi Wakabayashi, Mikio Shoji, and Hidenao Sasaki. Microrna expression profiles of neuron-derived extracellular vesicles in plasma from patients with amyotrophic lateral sclerosis. *Neuroscience letters*, 2019.

[37] Jaekwang Kim, Hyejin Yoon, Cristina M Ram´ırez, Sang-Mi Lee, Hyang-Sook Hoe, Carlos Fern´andezHernando, and Jungsu Kim. Mir-106b impairs cholesterol efflux and increases aβ levels by repressing abca1 expression. *Experimental neurology*, 235(2):476–483, 2012.

[38] Mariana Saint Just Ribeiro, Magnus L Hansson, Mikael J Lindberg, Anita E Popko-Scibor, and Annika E´ Wallberg. Gsk3β is a negative regulator of the transcriptional coactivator maml1. *Nucleic acids research*, 37(20):6691–6700, 2009.

[39] Maiko Moriyama, Takeshi Fukuhara, Markus Britschgi, Yingbo He, Ramya Narasimhan, Saul Villeda, Hector Molina, Brigitte T Huber, Mike Holers, and Tony Wyss-Coray. Complement receptor 2 is expressed in neural progenitor cells and regulates adult hippocampal neurogenesis. *Journal of Neuroscience*, 31(11):3981–3989, 2011.

[40] Yifeng Du, Michael C Wooten, and Marie W Wooten. Oxidative damage to the promoter region of sqstm1/p62 is common to neurodegenerative disease. *Neurobiology of disease*, 35(2):302–310, 2009.

[41] Yvonne Tay, Lev Kats, Leonardo Salmena, Dror Weiss, Shen Mynn Tan, Ugo Ala, Florian Karreth, Laura Poliseno, Paolo Provero, Ferdinando Di Cunto, et al. Coding-independent regulation of the tumor suppressor pten by competing endogenous mrnas. *Cell*, 147(2):344–357, 2011.

[42] Valerio Costa, Roberta Esposito, Marianna Aprile, and Alfredo Ciccodicola. Non-coding rna and pseudogenes in neurodegenerative diseases:"the (un) usual suspects". *Frontiers in genetics*, 3:231, 2012.

Figure legend:

Figure 1: The heterogeneous mutation data and the pipeline of GDAMDB model. (a. The idea of bridging heterogeneous *mutation type* and *mutation association* by using data synchronization. b. Graphical model of the "Mutation Data Bridging" model. c. pipeline of GDAMDB model for gene-disease association prediction.)

Figure 2: The illustrated examples of *mutation types* retrieved from Glu693Gly, Ala673Val, and Val717Ile/Cly/Phe in APP.

Figure 3: Biological process categories of 325 mutation types and the sentence evidences.

Figure 4: Data fusion of heterogeneous AD mutation data improves the discovery of novel AD-related genes.

Figure 5: 69 out of 79 predicted genes have supportive AD-related evidence.

Figure 6: The parameters setting of "Mutation Data Bridging" model.

Figures

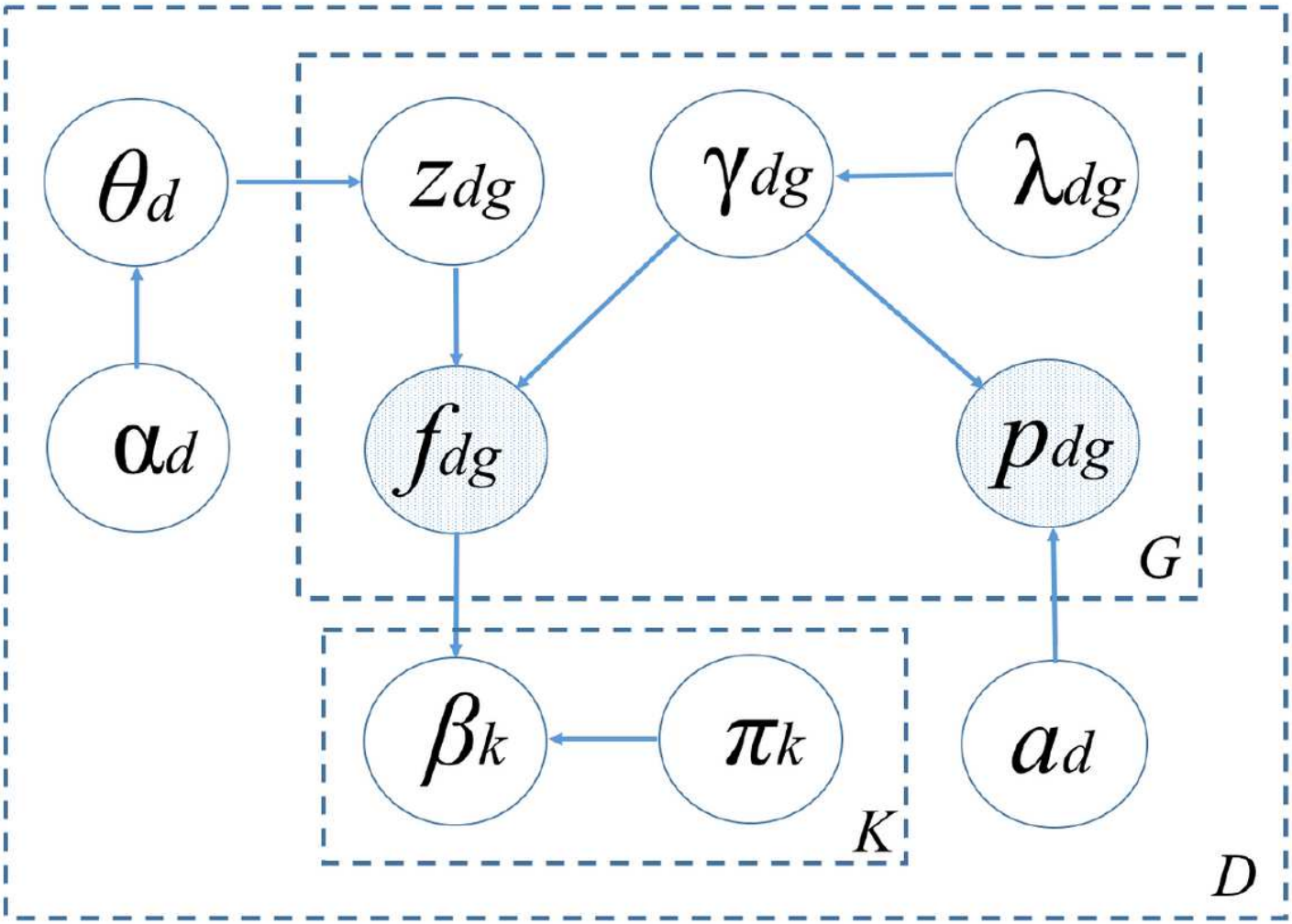

Figure 1

The parameters setting of “Mutation Data Bridging” model.

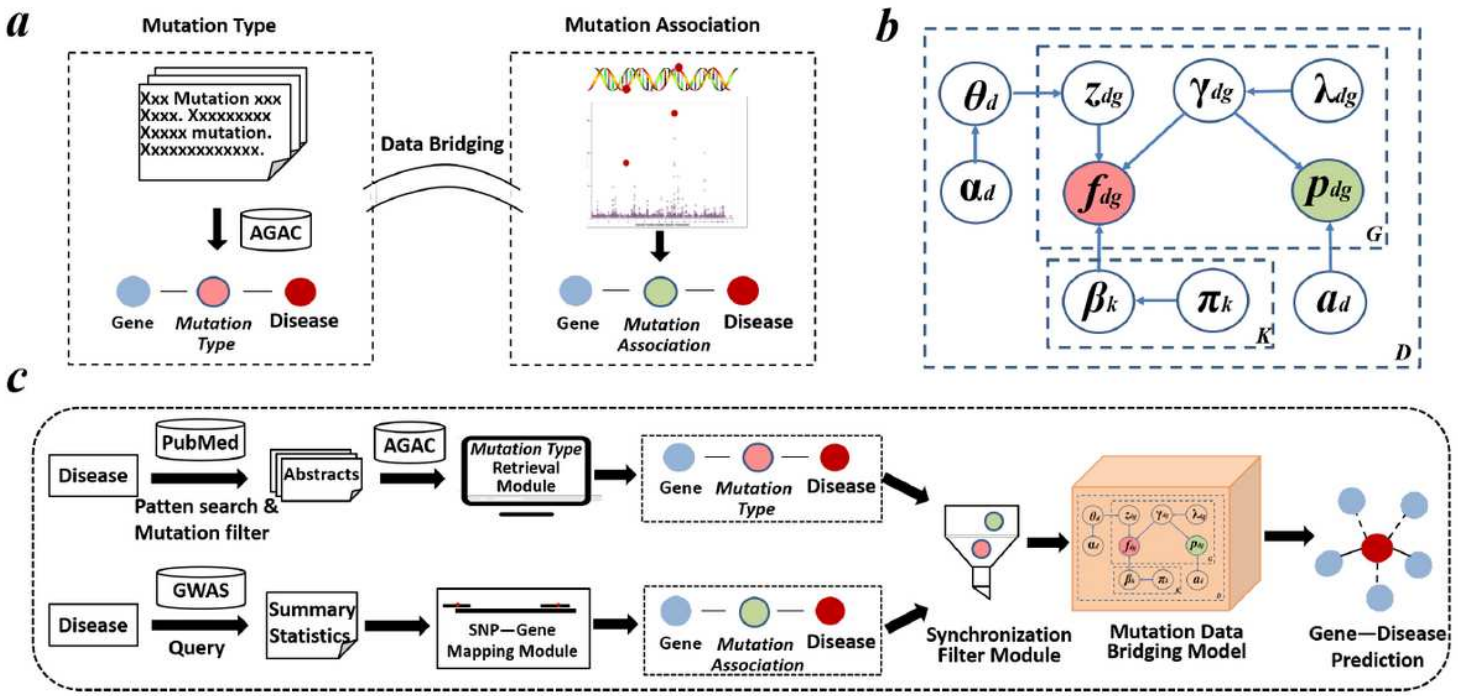

Figure 2

The heterogeneous mutation data and the pipeline of GDAMDB model. (a. The idea of bridging heterogeneous mutation type and mutation association by using data synchronization. b. Graphical model of the “Mutation Data Bridging” model. c. pipeline of GDAMDB model for gene-disease association prediction.)

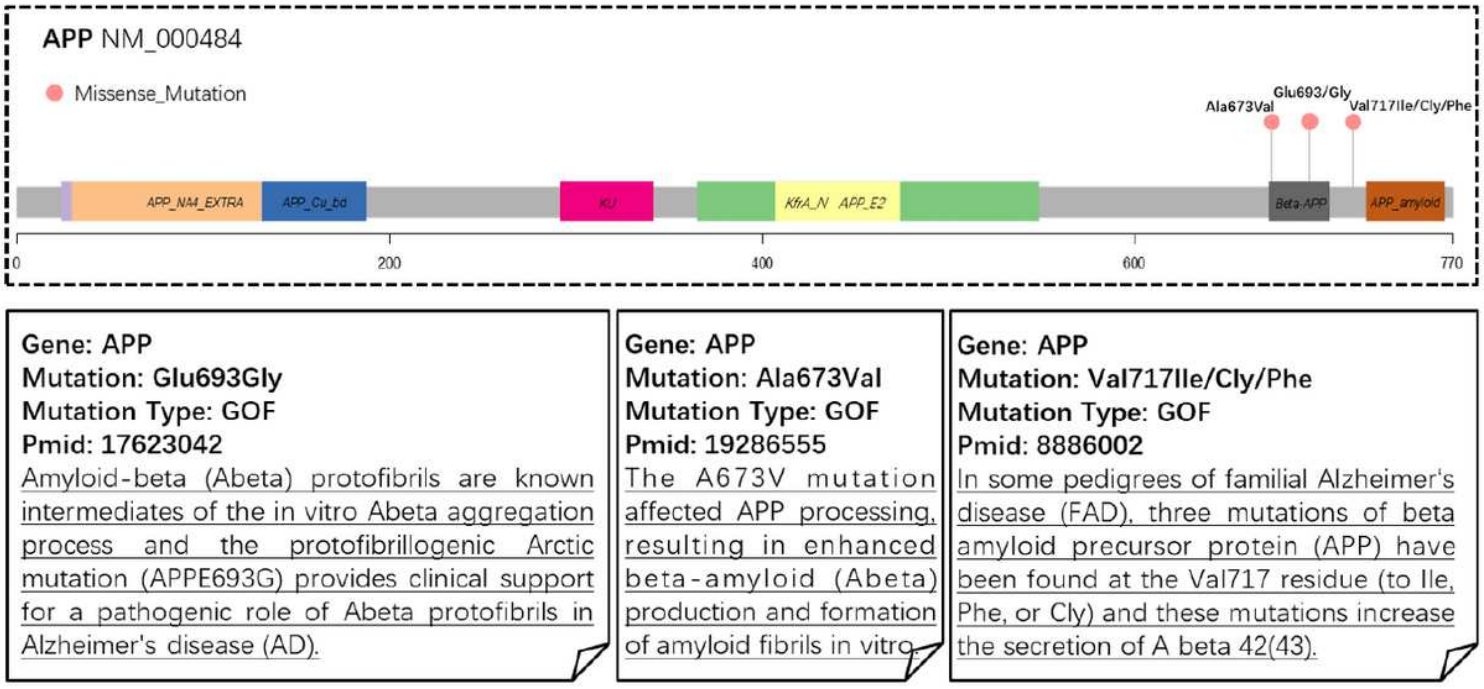

Figure 3

The illustrated examples of mutation types retrieved from Glu693Gly, Ala673Val, and Val717Ile/Cly/Phe in APP.

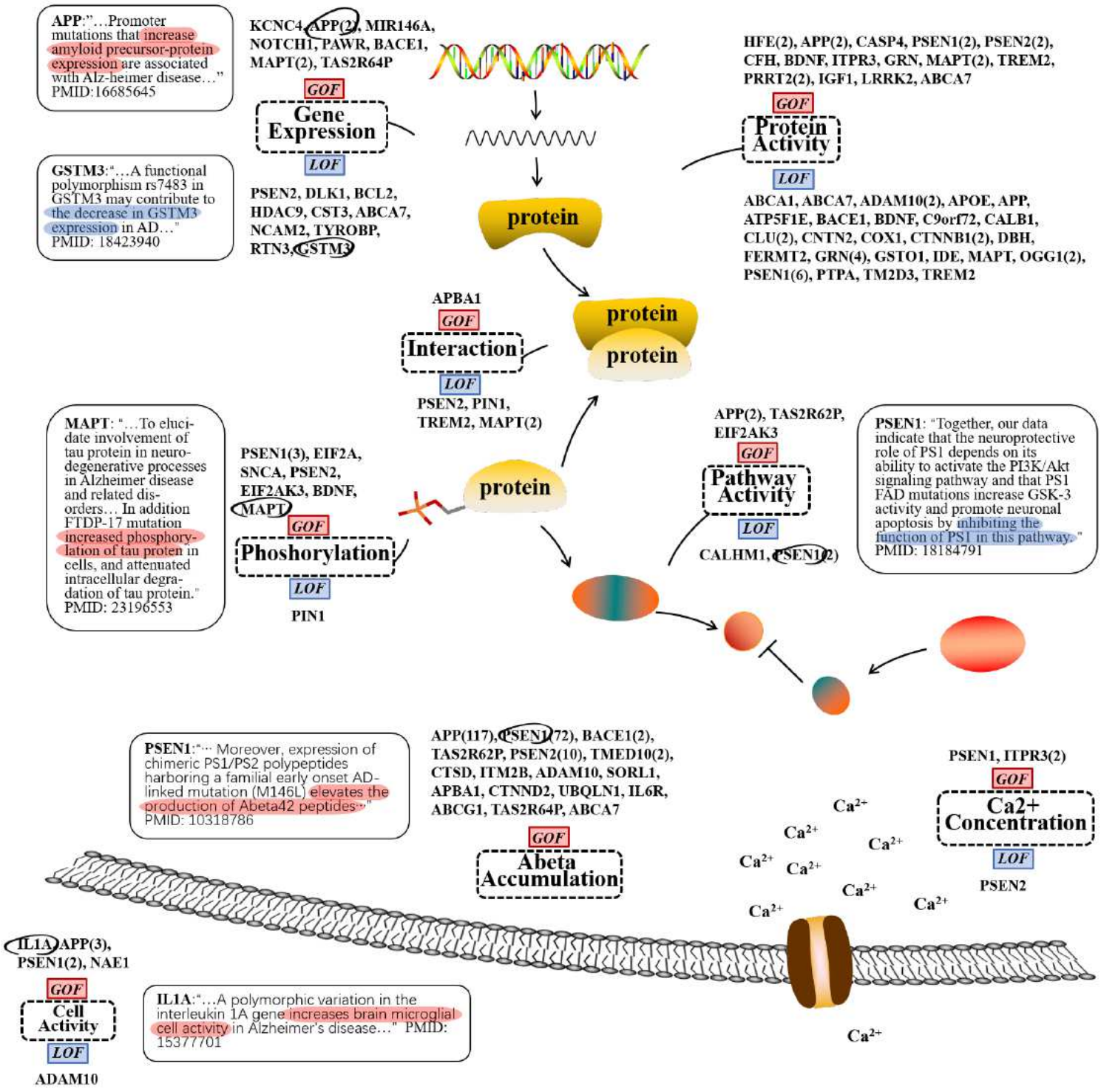

Figure 4

Biological process categories of 325 mutation types and the sentence evidences.

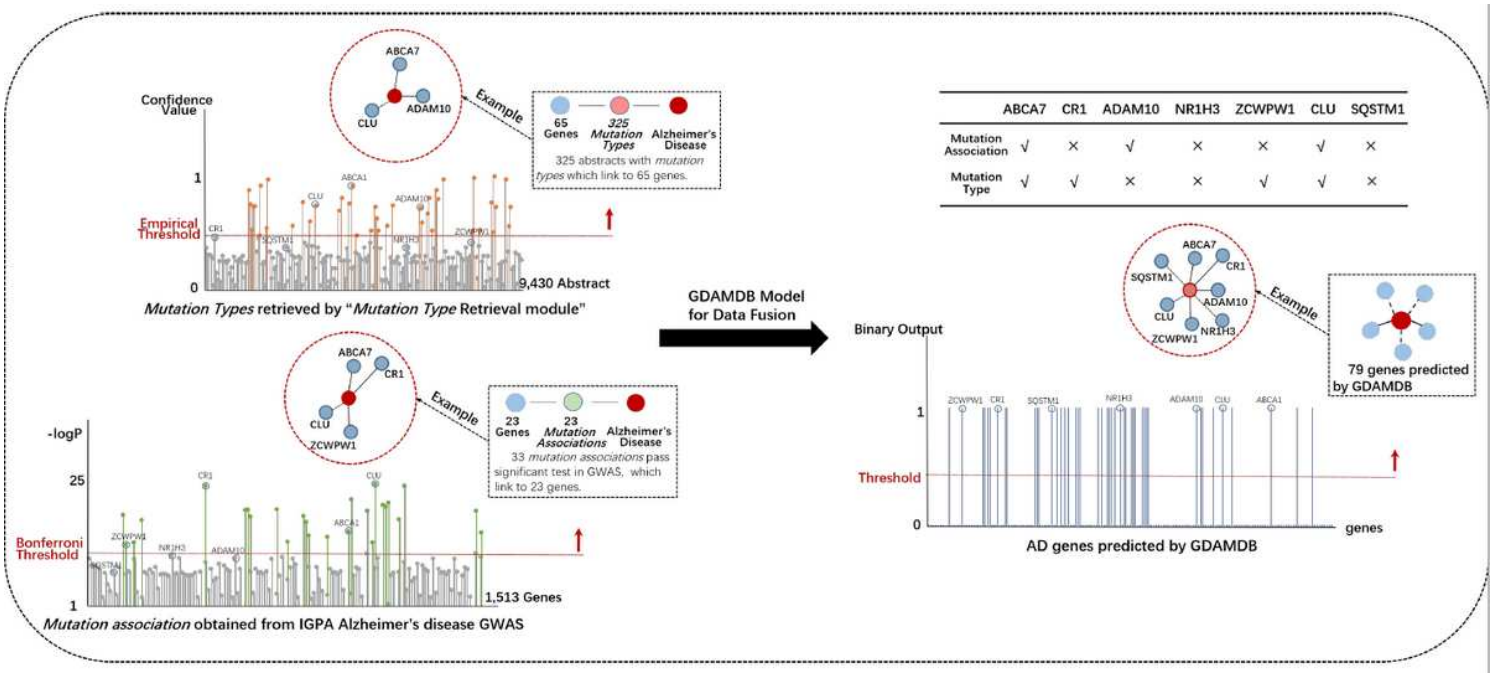

Figure 5

Data fusion of heterogeneous AD mutation data improves the discovery of novel AD-related genes.

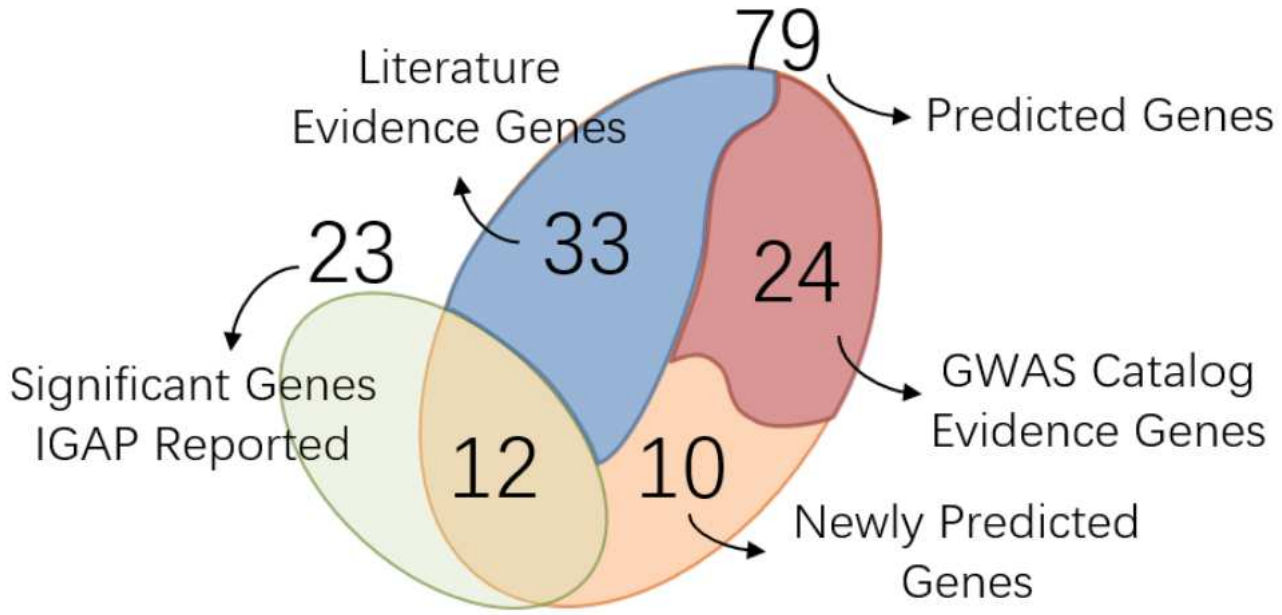

Figure 6

69 out of 79 predicted genes have supportive AD-related evidence.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S379PredictedGenes.xlsx
- S2MutationTypeData.xlsx
- S1ProofofMutationDataBridgingModel.pdf