

# Potential *Arabidopsis thaliana* glucosinolate genes identified from the co-expression modules using graph clustering approach

Sarahani Harun<sup>1</sup>, Nor Afiqah-Aleng<sup>2</sup>, Mohammad Bozlul Karim<sup>3</sup>, Md Altaf Ul Amin<sup>3</sup>, Shigehiko Kanaya<sup>3</sup> and Zeti-Azura Mohamed-Hussein<sup>1,4</sup>

<sup>1</sup> Centre for Bioinformatics Research, Institute of Systems Biology (INBIOSIS), Universiti Kebangsaan Malaysia, UKM Bangi, Selangor, Malaysia

<sup>2</sup> Institute of Marine Biotechnology, Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia

<sup>3</sup> Graduate School of Science and Technology & NAIST Data Science Center, Nara Institute of Science and Technology, Nara, Japan

<sup>4</sup> Department of Applied Physics, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, UKM Bangi, Selangor, Malaysia

## ABSTRACT

**Background:** Glucosinolates (GSLs) are plant secondary metabolites that contain nitrogen-containing compounds. They are important in the plant defense system and known to provide protection against cancer in humans. Currently, increasing the amount of data generated from various omics technologies serves as a hotspot for new gene discovery. However, sometimes sequence similarity searching approach is not sufficiently effective to find these genes; hence, we adapted a network clustering approach to search for potential GSLs genes from the *Arabidopsis thaliana* co-expression dataset.

**Methods:** We used known GSL genes to construct a comprehensive GSL co-expression network. This network was analyzed with the DPclusOST algorithm using a density of 0.5, 0.6, 0.7, 0.8, and 0.9. Generating clusters were evaluated using Fisher's exact test to identify GSL gene co-expression clusters. A significance score (SScore) was calculated for each gene based on the generated *p*-value of Fisher's exact test. SScore was used to perform a receiver operating characteristic (ROC) study to classify possible GSL genes using the ROCR package. ROCR was used in determining the AUC that measured the suitable density value of the cluster for further analysis. Finally, pathway enrichment analysis was conducted using ClueGO to identify significant pathways associated with the GSL clusters.

**Results:** The density value of 0.8 showed the highest area under the curve (AUC) leading to the selection of thirteen potential GSL genes from the top six significant clusters that include *IMDH3*, *MVP1*, *T19K24.17*, *MRSA2*, *SIR*, *ASP4*, *MTO1*, *At1g21440*, *HMT3*, *At3g47420*, *PS1*, *SAL1*, and *At3g14220*. A total of Four potential genes (*MTO1*, *SIR*, *SAL1*, and *IMDH3*) were identified from the pathway enrichment analysis on the significant clusters. These genes are directly related to GSL-associated pathways such as sulfur metabolism and valine, leucine, and isoleucine biosynthesis.

Submitted 6 May 2021  
Accepted 6 July 2021  
Published 4 August 2021

Corresponding author  
Zeti-Azura Mohamed-Hussein,  
zeti.hussein@ukm.edu.my

Academic editor  
Gerrit Beemster

Additional Information and  
Declarations can be found on  
page 18

DOI 10.7717/peerj.11876

© Copyright  
2021 Harun et al.

Distributed under  
Creative Commons CC-BY 4.0

## OPEN ACCESS

This approach demonstrates the ability of the network clustering approach in identifying potential GSL genes which cannot be found from the standard similarity search.

**Subjects** Bioinformatics, Computational Biology, Plant Science

**Keywords** Secondary metabolites, Nitrogen-containing compounds, Aliphatic glucosinolates, Indolic glucosinolates, Graph clustering, Gene network analysis

## INTRODUCTION

Plant secondary metabolites are divided into three chemically distinct classes: terpenes, phenolics, and nitrogen-containing compounds (Taiz & Zeiger, 2010). Alkaloids, cyanogenic glycosides, non-protein amino acids (NPAAs), and glucosinolates are the examples of nitrogen-containing compounds (Wink, 2015). Glucosinolates (GSLs) are known for protecting plants against invading pests and pathogens as well as preventing cancer in humans (Tang et al., 2010; Lai et al., 2010; Frerigmann et al., 2016; Megna et al., 2016; Piślewska-Bednarek et al., 2017). GSLs are found in the Brassicaceae family known as cruciferous vegetables, consisting of broccoli, cabbage, cauliflower, kale, mustard and cress (Herr & Büchler, 2010), and in the model plant, *Arabidopsis thaliana* (Redovniković et al., 2008). In 2001, 34 GSLs were identified from the leaves and seeds of 39 different *Arabidopsis* ecotypes, most of which were chain-elongated products originated from methionine (Met) (Kliebenstein et al., 2001).

GSLs are one of the most studied secondary metabolites since the beginning of 2000 until now (Sønderby, Geu-flores & Halkier, 2010; Burow et al., 2015). Interests in plant GSLs rise in recent years because of their importance in plant defense and cancer preventive agent (Tang et al., 2011; Megna et al., 2016), and other beneficial effects such as providing regulatory function inflammation, stress responses, antioxidant and antimicrobial properties (Bischoff, 2016). GSLs are amino acid-derived compounds divided into three main categories: aliphatic GSLs, derived from Ala, Leu, Ile, Val, and Met; benzyl GSLs, derived from Phe or Tyr; and indolic GSLs, derived from Trp. The GSL pathway consists of several genes that encode for transcription factors, transporters and enzymes involved in the biosynthesis of GSLs. The genes are essential in the side-chain elongation, core structure synthesis, side-chain modification, as well as GSL degradation (Agerbirk & Olsen, 2012; Blažević et al., 2019).

In recent years, the identification of genes involved in the GSL biosynthesis has been extensively investigated using gene co-expression data (Gachon et al., 2005; Hirai et al., 2005, 2007; Knill et al., 2008; Sawada et al., 2009a). This approach is used to identify novel genes that encode for transcription factors (TFs) and enzymes involved in the GSL biosynthesis (Hirai et al., 2005, 2007; Knill et al., 2008; Sawada et al., 2009a). Hirai et al. (2007) have proven the association between MYB28 and MYB29 with aliphatic GSL biosynthesis; both genes were previously unknown to be responsible in encoding R2R3-Myb TFs. Their analysis has also shown the co-expression of TFs with several known aliphatic GSLs: cytochrome P450 79F1 (CYP79F1), cytochrome P450 79F2

(*CYP79F2*), methylthioalkylmalate synthase 1 (*MAM1*) and methylthioalkylmalate synthase 3 (*MAM3*) (*Hirai et al., 2007*). Meanwhile, *Sawada et al. (2009b)* have identified bile acid transporter 5 (*BAT5*) to co-express with the aliphatic GSL genes involved in chain elongation.

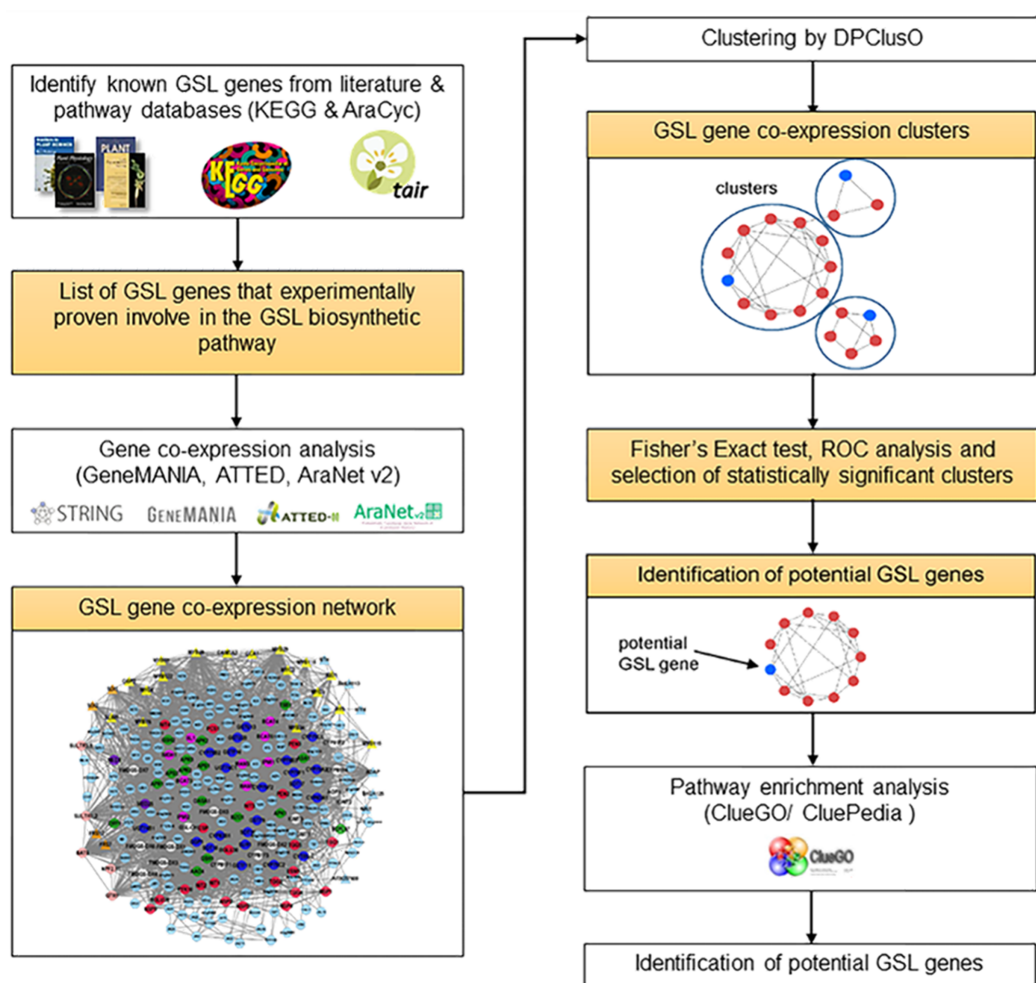
Recently (*Harun et al., 2020*) have compiled a total of 113 known GSL genes with experimental evidences from the published research conducted for the last 20 years. They classified the genes according to their annotation and grouped them into TFs, biosynthetic genes, and protein transporters. These genes are used as bait genes to find “additional/missing genes” from the co-expression modules in order to identify more novel GSL genes. Thus, a computational pipeline of biological network approach can accelerate the finding of these genes. Previous studies demonstrated the application of the graph clustering approach in the protein-protein interaction (PPI) network followed by the Fisher’s exact test to identify disease clusters in the inflammatory bowel disease (IBD) (*Eguchi et al., 2018*) and polycystic ovarian syndrome (PCOS) (*Afiqah-Aleng et al., 2020*). The identification of the novel disease genes and related pathways is able to discover disease-associated genes that are linked with other diseases as well. Hence, understanding the biological components involved would lead to additional insight into the mechanism of both diseases that lead to effective treatment in the future (*Eguchi et al., 2018*; *Afiqah-Aleng et al., 2020*). In *Oryza sativa*, a clustering approach of a PPI network was able to elucidate a molecular mechanism of nitrate that regulates nitrite reductase, ferredoxin-NADP reductase, and ferredoxin. These three components are associated with flowering time that led to a novel contribution of nitrate signaling in light and dark environment (*Pathak et al., 2020*). In this article, we describe the computational pipeline that involves the use of graph clustering for novel genes annotation in plants. In this study, we used the abovementioned approach to search for potential GSLs biosynthetic genes which were previously unknown genes.

## MATERIALS & METHODS

### Data collection and co-expression network construction

Before this study, only 46 known GSL genes were identified by *Sønderby, Geu-Flores & Halkier (2010)*. Following that, we performed a comprehensive literature and pathway databases search in KEGG (<http://www.genome.jp/kegg/>) (*Kanehisa et al., 2016*) and AraCyc (<https://www.arabidopsis.org/biocyc/>) (*Mueller, Zhang & Rhee, 2003*) to identify more known GSL genes using keywords, such as glucosinolate AND Arabidopsis, in journals published in 2020 (*Harun et al., 2020*). The list of updated GSL genes were also added in our manually curated sulfur compound database, SuCComBase (<http://plant-scc.org/>) (*Harun et al., 2019*).

We utilized ATTED-II database (<http://atted.jp/>) (*Aoki et al., 2016*), a database of co-expressed genes, initially involving *Arabidopsis* and rice, to identify candidate genes co-expressed with known GSLs. In ATTED, the *Arabidopsis* RNAseq and microarray data from ArrayExpress (*Rustici et al., 2013*) covered 94% and 76% of the *Arabidopsis* protein-encoding genes, respectively. In this study, we used three additional co-expression tools: AraNet v2 (*Lee et al., 2014*), GeneMANIA (*Lee et al., 2014*), and



**Figure 1** Step-by-step procedure to identify potential GSL genes involved in the GSL biosynthetic pathway. Full-size [DOI: 10.7717/peerj.11876/fig-1](https://doi.org/10.7717/peerj.11876/fig-1)

STRING (Szklarczyk et al., 2015, 2017, 2019). The co-expression data in AraNet covers 83.5% of the *Arabidopsis* coding genome from the Gene Expression Omnibus (GEO) database involving 1,261 microarray experiments (Barrett et al., 2013).

We used known GSL genes as a query against three co-expression tools in an effort to search for “additional” genes that were co-expressed with them. We defined “additional” genes as potential GSL genes that will be critically and systematically assessed using cluster and pathway enrichment analysis before mapping them on the GSL biosynthesis pathway. Known GSL genes were used as a query against the co-expression tools, including ATTED-II version 10.1 (<http://atted.jp/>) (Aoki et al., 2016), AraNet v2 (Lee et al., 2014), GeneMANIA (Warde-Farley et al., 2010; Montojo et al., 2014) and STRING (Szklarczyk et al., 2015, 2017, 2019). A total of Four gene networks were combined into a single gene co-expression network using Cytoscape 3.7.1 (Shannon et al., 2003). The steps involved in data establishment and gene co-expression network construction are shown in Fig. 1.



The list of plant transcription factors (TFs) in *A. thaliana* was downloaded from the Plant Transcription Factor Database v5.0 (PlantTFDB 5.0; <http://planttfdb.cbi.pku.edu.cn>) (Jin *et al.*, 2016). The information generated from PlantTFDB will add Biological information of each potential GSL gene was added in this study using various databases, such as UniProt (<https://www.uniprot.org/>) (Bateman *et al.*, 2017) and TAIR (<https://www.arabidopsis.org/index.jsp>) (Lamesch *et al.*, 2012).

### Calculating clusters

DPCLUSOST (Bozlul Karim, Wakamatsu & Altaf-Ul-Amin, 2017), an option in DPCLUSO algorithm (Altaf-Ul-Amin *et al.*, 2006; Altaf-Ul-Amin, Wada & Kanaya, 2012) was used to generate clusters in order to identify densely connected regions from a gene network using a graphical interface. The clustering algorithm generates overlapping clusters that influenced several biological processes related to GSL metabolism. DPCLUSO is used for an undirected graph consisting of a finite set of nodes  $N$  and a finite set of edges  $E$ . In this algorithm, two critical parameters are introduced: density  $d_k$  and cluster property  $cp_{nk}$ . Density  $d_k$  of cluster  $k$  refers to the ratio of the number of actual cluster edges ( $|E_k|$ ) and the maximum possible number of cluster edges ( $|E_k|_{\max}$ ). Detailed information on this algorithm was described in previous studies by Eguchi *et al.* (2018) and Afiqah-Aleng *et al.* (2020). The cluster property of node  $n$  with respect to cluster  $k$  is represented by the following equation:

$$cp_{nk} = \frac{|E_{nk}|}{d_k \times |N_k|} \quad (1)$$

$N_k$  refers to the number of nodes in cluster  $k$ .  $E_{nk}$  is the total number of edges connecting the node  $n$  with nodes of cluster  $k$ .

### Fisher's exact test

Fisher's exact test (Fisher, 1992) was used to evaluate known GSL gene enrichment clusters. It is a statistical test used in the analysis of  $2 \times 2$  contingency tables (Fisher, 1922, 1992). The introduced values of  $a$ ,  $b$ ,  $c$  and  $d$  are shown in Table 1. In order to identify the best set of clusters, a calculation to obtain the average significance of a cluster set was introduced. Fisher's exact test  $p$ -values were calculated to assess GSL genes enrichment in each of the identified clusters.

### SScore and ROC analysis

The prediction confidence of potential GSL genes was calculated for each gene depending on the  $p$ -value of the generated clusters using a significance score (SScore). A similar approach was used in the protein-protein interaction network on human diseases as described by Eguchi *et al.* (2018) and Afiqah-Aleng *et al.* (2020). The formula for SScore was  $SScore = -\log(p\text{-value})$ . Since DPCLUSO produced overlapping clusters, the lowest  $p$ -value of a gene was used to measure SScore. A gene can belong to more than one cluster and equate to more than one  $p$ -value. Next is the receiver operating characteristic (ROC) analysis that was conducted to identify the potential GSL genes by calculating the power of

**Table 1** The contingency table prepared in this study to calculate known GSL gene enrichment clusters.

|                | GSL genes    | Non-GSL genes |
|----------------|--------------|---------------|
| In cluster     | <i>a</i>     | <i>b</i>      |
| Not in cluster | <i>c</i>     | <i>d</i>      |
|                | <i>a + c</i> | <i>b + d</i>  |

**Note:**

<sup>a</sup>Here *n* refer to the total number of genes in the gene network.

SScore (Metz, 1978; Davis & Goadrich, 2006). True Positive Rate (TPR) and False Positive Rate (FPR) were calculated using a series of threshold (*th*) SScore in this study.

The fraction of true positive predictions in all positive data is TPR, and the fraction of false positive predictions in all negative data is FPR. The following equations were used to calculate TPR and FPR:

$$TPR = \frac{TP}{TP + FN} \quad (2)$$

$$FPR = \frac{FP}{FP + TN} \quad (3)$$

Based on the listed equations above, true positive (TP), false positive (FP), true negative (TN), and false negative (FN) were known as the number of GSL genes with  $SScore \geq th$ , number of non-GSL genes with  $SScore \geq th$ , number of non-GSL genes having  $SScore < th$  and number of GSL genes having  $SScore < th$ , respectively. The Area Under the ROC Curve (AUC) study was used to evaluate the efficiency of SScore in identifying potential GSL genes. ROCR (Sing et al., 2005), a R package, was used to calculate the AUC in this study.

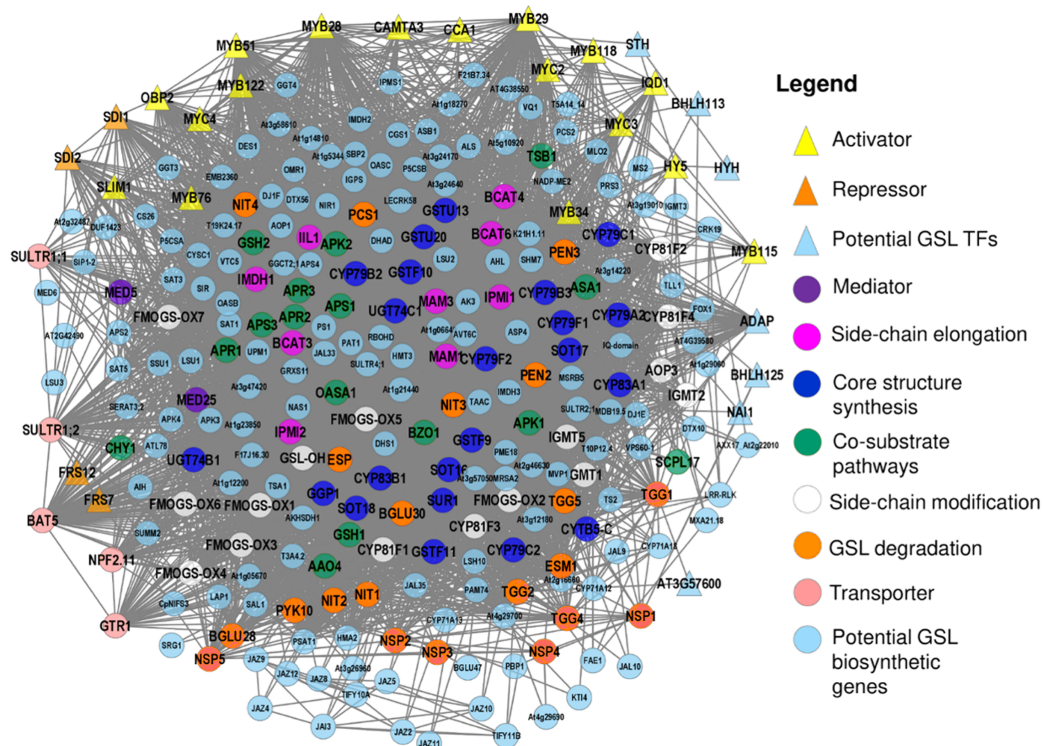
### Pathway enrichment analysis

To evaluate the biological role of the clusters, pathway enrichment analysis was performed on the potential GSL genes and GSL clusters against pathway databases, including Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al., 2017) using ClueGO/CluePedia (Bindea et al., 2009) apps in Cytoscape. The false discovery rate of each pathway was calculated using a hypergeometric test with Bonferroni correction to determine its importance. To define the relation between pathways, a Kappa score of 0.5 was chosen. The overview of each step taken in the gene network clustering approach, statistical analysis on the significant clusters, and pathway enrichment analysis are shown in Fig. 1.

## RESULTS

### Identification of GSL genes and co-expression network construction

All information on the genes encoding proteins involved in GSL was extracted using the sources shown in Fig. 1. Unlike Kyoto Encyclopedia of Genes and Genomes (KEGG), there



**Figure 2** A gene co-expression network that consists of 250 nodes and 5,554 edges.

Full-size [DOI: 10.7717/peerj.11876/fig-2](https://doi.org/10.7717/peerj.11876/fig-2)

are 12 GSL pathway derivatives in AraCyc: aliphatic GSL biosynthesis (side-chain elongation cycle), GSL biosynthesis from homomethionine, GSL biosynthesis from dihomomethionine, GSL biosynthesis from trihomomethionine, GSL biosynthesis from tetrahomomethionine, GSL biosynthesis from pentahomomethionine, GSL biosynthesis from hexahomomethionine, GSL biosynthesis from phenylalanine, GSL biosynthesis from tryptophan, GSL breakdown, indole GSL breakdown (active in intact plant cell) and indole GSL breakdown (insect chewing induced). Finally, a total of 113 known GSL genes (experimentally verified GSL genes) were used throughout this study.

The 113 known GSL genes were used as bait genes to query the whole transcriptomics data from four co-expression network tools (ATTED, GeneMANIA, STRING, and AraNet v2). **Figure 2** shows the interaction between 112 GSL genes with 158 interacting partners, generating 5,554 edges. This network was constructed from four gene co-expression networks that produced 161 nodes and 4,108 edges from GeneMANIA; 88 nodes and 355 edges from AraNet; 161 nodes and 2325 edges from STRING, and 197 nodes and 370 edges from ATTED. These individual gene co-expression networks were merged using Cytoscape to produce an integrated gene co-expression network consisting of 270 nodes and 5,554 edges (**Fig. 2**).

The gene co-expression network (**Fig. 2**) consists of various functional groups based on the mechanism in GSL biosynthesis. *AOP2* (alkenyl hydroxyalkyl producing 2) was the only known GSL gene that was not found in this gene network. The absence of *AOP2* in the

**Table 2** Cluster properties of different input densities using DPCLUSO algorithm.

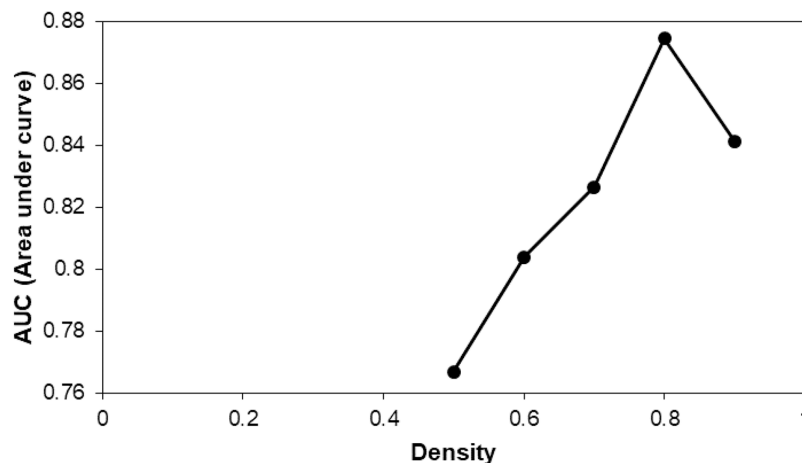
| Density | Number of cluster | maxsize | avgsize |
|---------|-------------------|---------|---------|
| 0.5     | 95                | 125     | 31.41   |
| 0.6     | 152               | 102     | 17.13   |
| 0.7     | 163               | 82      | 62.96   |
| 0.8     | 186               | 64      | 42.44   |
| 0.9     | 213               | 47      | 15.08   |

gene network might due to lack of co-expression data that link the bait gene with other genes in the co-expression databases. The transcriptional components in GSL biosynthesis were divided into their respective GSL regulatory mechanism, such as activator, repressor, and mediator. The GSL regulatory network would affect several GSL biosynthetic pathways that can be grouped into side-chain elongation, core structure synthesis, co-substrate pathways, and side chain modification (Harun et al., 2020). GSL degradation refers to the formation of activated GSL products that are known to confer protection in plants against the biotic, and abiotic stresses (Halkier & Gershenzon, 2006; Liu et al., 2020). There are also five known GSL transporters in the gene network: BAT5, SULTR1;1, SULTR1;2, GTR1, and GTR2. In Fig. 2, the 158 additional genes that are defined as interacting partners with the known GSL are characterized into seven potential GSL transcription factors (TFs), and 151 potential biosynthetic GSL genes that will be further analyzed in this study.

### Gene co-expression clusters analysis

Once the gene co-expression network was constructed, the present clusters in the network were determined using the DPCLUSOST algorithm. DPCLUSOST extracts highly interconnected region that perform a similar biological process. We hypothesize that co-exist genes with known GSL genes in the same statistically significant clusters can be used to predict potential GSL genes. These co-exist genes are the additional genes in the gene co-expression network clustered with known GSL genes. A total of five sets of clusters were generated using density values of 0.5, 0.6, 0.7, 0.8, and 0.9 with  $cp$  value of 0.5 (Table 2). The density  $d_k$  of any cluster  $k$  refers to the ratio of the number of edges in the cluster ( $|E_k|$ ) and the maximum possible number of cluster edges ( $|E_k|_{\max}$ ). Clusters generated from different density values produced distinctive cluster characteristics, namely the number of clusters, the maximum size of the cluster, and the average cluster size. Smaller density values resulted in greater cluster sizes and fewer clusters, as expected. As for  $cp$  value, 0.5 is the default and recommended value and has been used in previous studies (Eguchi et al., 2018; Karim et al., 2020).

From the five different input densities, DPCLUSOST generated five sets of clusters. To determine which set of clusters for further analysis, we performed a receiver operating characteristic (ROC) analysis. First, Fisher's exact test  $p$ -values were calculated to assess GSL genes enrichment in each of the identified clusters. Then we assigned the significance score (SScore), to each gene, based on the  $p$ -values of the clusters to which they belong.



**Figure 3** The calculated area under curve (AUC) of five clusters generated using DPCLUSO. The density value of 0.8 has the highest AUC, followed by clusters generated from density value of 0.9, 0.7, 0.6, and 0.5. [Full-size !\[\]\(b345a1c4255362eec3746050dd71ccac\_img.jpg\) DOI: 10.7717/peerj.11876/fig-3](https://doi.org/10.7717/peerj.11876/fig-3)

Next, we created five ROC curves by utilizing the SScore corresponding to the five sets of clusters. The AUC of five ROC curves is shown in Fig. 3. The maximum AUC was 0.87, generated from the density value of 0.8. The potential GSL genes found within the statistically significant clusters of the set corresponding to density 0.8 were selected as potential GSL genes (Table S1).

A total of 148 significant clusters with a density value of 0.8 ( $p$ -value < 0.05) was identified, with the potential GSL genes found within the statistically significant cluster being considered significant and analyzed further in this study. The overall result of the 148 significant clusters are shown in Table S2. Table 3 shows a list of potential GSL genes identified from the selected highly significant clusters. Based on Table S2, the top six highly significant clusters (Cluster 121, Cluster 131, Cluster 127, Cluster 125, Cluster 129, and Cluster 128) were chosen for further analysis. The genes in the light blue nodes referred to the potential GSL genes in our study.

Based on Table 3, a total of thirteen potential GSL genes were identified from the top six significant clusters: *IMDH3*, *MVP1*, *T19K24.17*, *MRSA2*, *SIR*, *ASP4*, *MTO1*, *At1g21440*, *HMT3*, *At3g47420*, *PS1*, *SAL1*, and *At3g14220*. Each cluster contained known GSL genes with functions that included transcription factors and other related GSL biological processes, such as side-chain elongation, core structure synthesis, side-chain modification, and GSL degradation. These genes were also grouped into aliphatic and indolic GSL genes depending on their involvement in the type of GSL being produced. There were also a group of gene-encoding enzymes involved in GSL degradation and a GSL transporter (BAT5).

### Pathway enrichment analysis

KEGG pathway enrichment was identified from the selected top six significant clusters and five significant pathways, *i.e.*, selenocompound metabolism, sulfur metabolism, tryptophan metabolism, valine, leucine, and isoleucine biosynthesis, and GSL biosynthesis



**Table 3** List of potential GSL genes from selected highly significant clusters.

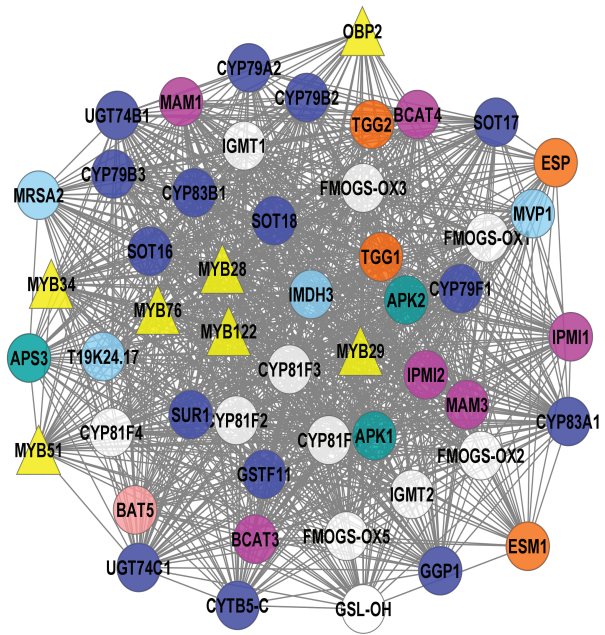
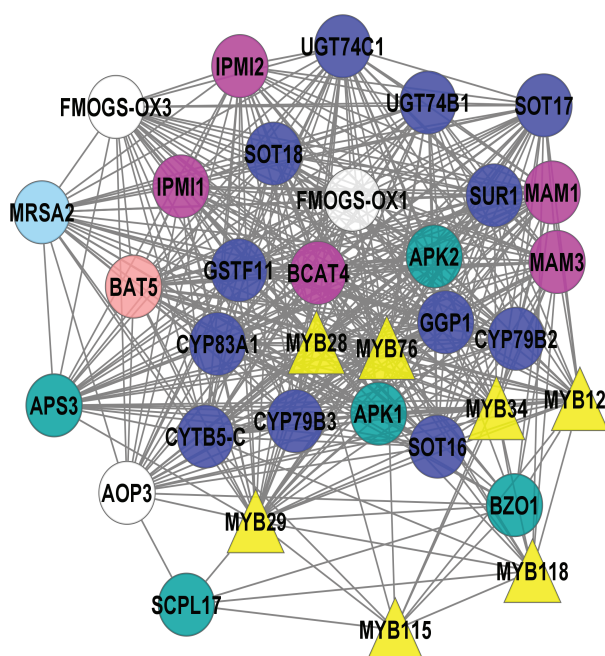
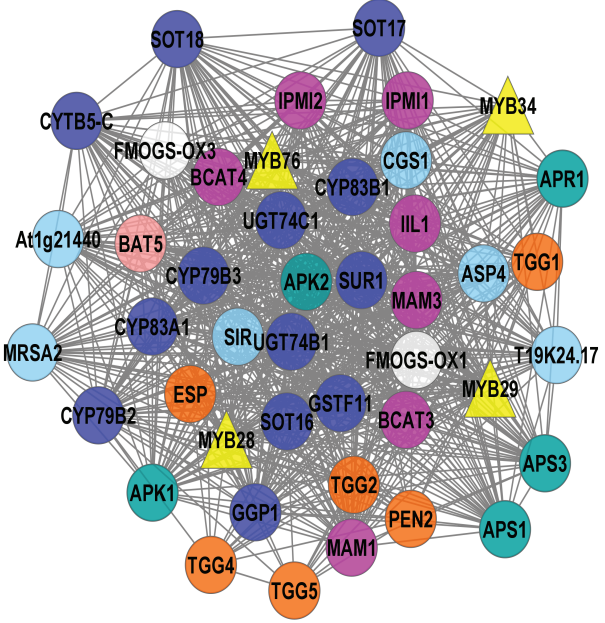
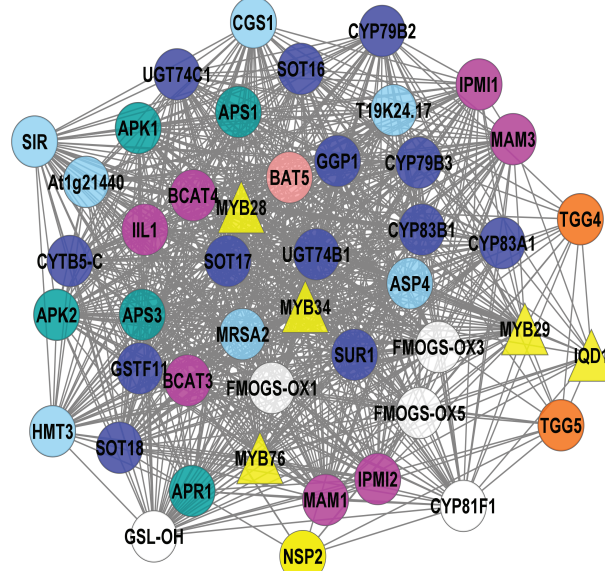
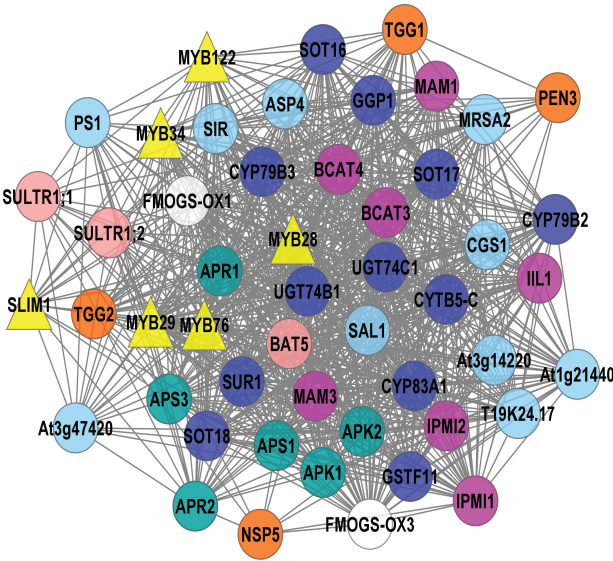
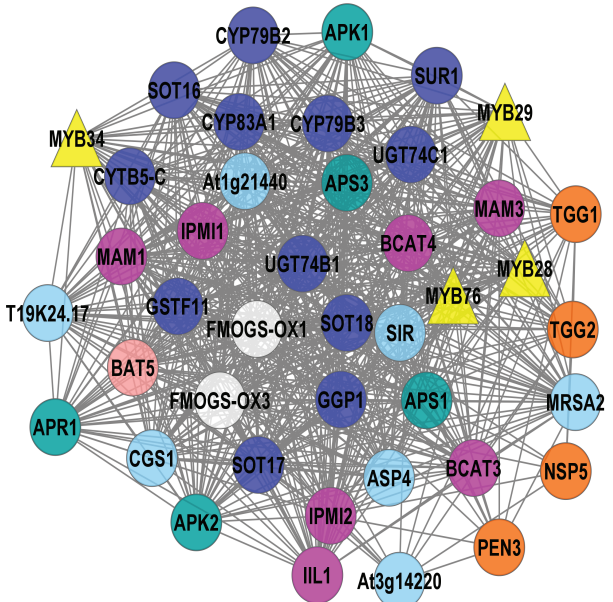
| Cluster number | Cluster size | Potential GSL genes (number)                                     | Cluster  | <i>p</i> -value |
|----------------|--------------|--|--|-----------------|
| Cluster 121    | 51           | <i>IMDH3</i> , <i>MVP1</i> , <i>T19K24.17</i> , <i>MRSA2</i> (4) |    | 5.79E-17        |
| Cluster 131    | 34           | <i>MRSA2</i> (1)   |  | 2.27E-13        |

Table 3 (continued)

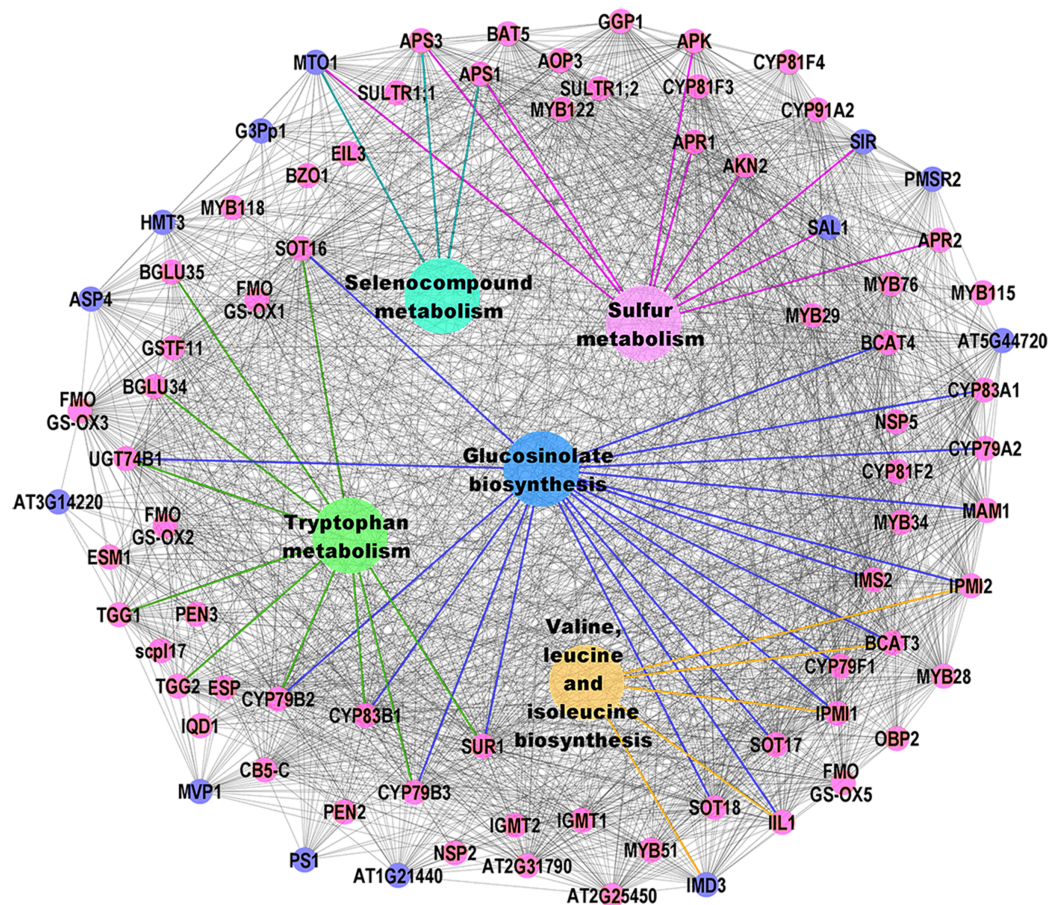
| Cluster number | Cluster size | Potential GSL genes (number)  | Cluster  | p-value  |
|----------------|--------------|---|--|----------|
| Cluster 127    | 44           | <i>T19K24.17</i> , <i>SIR</i> , <i>MRSA2</i> ,<br><i>ASP4</i> ,<br><i>CGS1</i> , <i>At1g21440</i> (6)               |   | 2.75E-11 |
| Cluster 125    | 46           | <i>T19K24.17</i> , <i>SIR</i> , <i>MRSA2</i> ,<br><i>ASP4</i> ,<br><i>CGS1</i> , <i>At1g21440</i> , <i>HMT3</i> (7) |  | 4.65E-11 |

(Continued)

Table 3 (continued)

| Cluster number | Cluster size | Potential GSL genes (number)   | Cluster  | <i>p</i> -value |
|----------------|--------------|--|--|-----------------|
| Cluster 129    | 41           | <i>At3g47420</i> , <i>PS1</i> , <i>SIR</i> , <i>T19K24.17</i> , <i>SAL1</i> , <i>CGS1</i> (6)                      |    | 4.77E-10        |
| Cluster 128    | 42           | <i>At3g14220</i> , <i>T19K24.17</i> , <i>SIR</i> , <i>MRSA2</i> , <i>ASP4</i> , <i>CGS1</i> , <i>At1g21440</i> (7) |  | 1.93E-09        |





**Figure 4** Visualization of KEGG pathway enrichment using ClueGO/CluePedia apps from Cytoscape. The enrichment shows only significant pathways ( $p$ -value  $\leq 0.05$ ).

Full-size [DOI: 10.7717/peerj.11876/fig-4](https://doi.org/10.7717/peerj.11876/fig-4)

(Fig. 4) were selected. Pink nodes denote the known GSL genes that encode for enzymes, TFs, and protein transporters whilst blue nodes denote the potential GSL genes might involve in GSL biosynthesis.

## DISCUSSION

In this study, we propose a combinatorial approach as an alternative way to identify possible GSL genes as compared to using the traditional sequence-based searching. Our method demonstrates that graph clustering analysis combined with Fisher's exact test and ROC analysis is able to classify possible genes that aren't identified or were missed using the standard sequence-based searching. We believe it is critical to provide an alternative technique that is able to search and classify the sequences without relying solely on the sequence searching technique. DPclusOST is a clustering algorithm that discovers and classifies strongly interconnected regions in a large network between core nodes, or high connectivity nodes and peripheral nodes, or low connectivity nodes to indicate biological significance in a cell. A similar approach was used in analysing protein-protein interaction network in inflammatory bowel disease (IBD) (Eguchi *et al.*, 2018) and

polycystic ovarian syndrome (PCOS) (Afiqah-Aleng *et al.*, 2020). DPCLUSOST creates overlapping clusters depending on a gene's multifunctionality, resulting in a high probability of a gene being present in multiple clusters (Altaf-UL-Amin, Wada & Kanaya, 2012; Bozlul Karim, Wakamatsu & Altaf-UL-Amin, 2017). The DPCLUSOST algorithm extracts highly interconnected region that perform similar biological process. Thus, integrating this algorithm in our pipeline demonstrate its ability to suggest that the existence of genes in the same statistically significant clusters with the known GSL genes can be used to predict potential GSL genes.

DPCLUSOST algorithm produced five sets of clusters generated from five density values. Fisher's exact test *p*-value and SScore with the AUC value were used to assess the cluster. The maximum AUC was identified from the clusters generated from density value of 0.8 producing 127 potential GSL genes (Table S1). One of the potential GSL gene, *ADAP* (ARIA-interacting double AP2-domain protein) was identified from the clustering approach and Fisher's exact test. Further functional analysis was performed using chimeric repressor gene silencing technology (CRES-T) and gene expression analysis using qPCR. The over-expression of downstream aliphatic GSL genes (*UGT74C1* and *IPMI1*) in the *ADAP*-SRDX line indicated the possibility of *ADAP* as a negative regulator in aliphatic GSL biosynthesis *via* a feedback mechanism (Harun *et al.*, 2021). A total of thirteen potential genes (Table 3) were identified from the top six significant analysis that is known as *IMDH3*, *MVP1*, *T19K24.17*, *MRSA2*, *SIR*, *ASP4*, *MTO1*, *At1g21440*, *HMT3*, *At3g47420*, *PS1*, *SAL1*, and *At3g14220*. Interestingly, *MRSA2* and *At1g21440* were found to be co-expressed with known aliphatic GSL biosynthesis in Arabidopsis in a large-scale analysis of plant gene co-expression network of specialized metabolic pathways. *MRSA2* is peptide methionine sulfoxide reductase protein whereas *At1g21440* is a phosphoenolpyruvate carboxylase family protein where they were both not known to be involved in the aliphatic GSL biosynthesis (Wisecaver *et al.*, 2017).

Pathway enrichment analysis using ClueGO/CluePedia (Bindea *et al.*, 2009) apps from Cytoscape (Shannon *et al.*, 2003) was used to interpret the association of the potential GSL genes and the selected significant clusters in Table 3. Enrichment analysis is used to map known biological functions of the generated clusters that were extracted from pathway databases such as KEGG (de Anda-Jáuregui, 2019). Figure 4 shows map of pathway enrichment-merged pathway that contains five enriched clusters that are related with GSL biosynthesis, *i.e.*, sulfur metabolism, tryptophan metabolism, and valine, leucine and isoleucine biosynthesis. Since GSLs have at least two sulfur atoms in their main structure, and aliphatic GSLs may have additional sulfur in their side chains, sulfur metabolism might have an importance to GSL biosynthesis (Falk, Tokuhisa & Gershenzon, 2007). Sulfur content in GSLs indicates that this compound is critical in GSL biosynthesis. Several metabolomic and transcriptomic studies reported a significant reduction in GSL accumulation under sulfur deficiency environment, suggesting the role of sulfur in GSL biosynthesis (Nikiforova *et al.*, 2003; Hirai *et al.*, 2003; Aarabi *et al.*, 2016). *SULTR1; 1* and *SULTR1; 2* in *Arabidopsis* roots play a role as sulfate transporters and their expression was increased in sulfur-limitation *Arabidopsis* (Koprivova & Kopriva, 2014; Morikawa-Ichinose *et al.*, 2019). The molecular components involving sulfate and GSL transport



machinery is more complex in Brassica crops, and requires an in-depth understanding on the GSL mechanism. From Fig. 4, several GSL genes are grouped in GSL core-substrate pathways, such as 5'-adenylylsulfate reductases (*APR1* and *APR2*), adenylyl-sulfate kinases (*APK1* and *APK2*), and ATP sulfurylases (*APS1* and *APS3*) and they were identified to be involved in both aliphatic and indolic GSL biosynthesis and linked to sulfate assimilation (Yatusevich et al., 2010a; Harun et al., 2020).

There are three potential GSL genes directly linked to sulfur metabolism (Fig. 4): cystathionine gamma-synthetase 1 (*MTO1*), sulfite reductase (*SIR*), and SAL1 phosphatase (*SAL1*). CGS is the key enzyme in methionine biosynthesis located in the chloroplast (Takahashi et al., 2011). *SIR* has been previously identified in sulfate assimilation that catalyzes the production of sulfide. Sulfide undergoes a cysteine biosynthesis as well as other sulfur-containing compounds, such as GSLs (Miao et al., 2016). Previous study showed significant increase in the *SIR* expression in *Arabidopsis* plants with overexpressed indolic GSL TFs (*MYB51*, *MYB122*, and *MYB34*) and aliphatic GSL TF (*MYB28*) (Yatusevich et al., 2010b). *SAL1* is a bifunctional enzyme that regulates the activities of 3' (2'),5'-bisphosphate nucleotidase and inositol polyphosphate 1-phosphatase (Quintero, Garcíadeblás & Rodríguez-Navarro, 1996). Ishiga et al. (2017) reported a reduced level of aliphatic GSLs production in the *sal1* mutants compared to the wild-type Col-0 in response to pathogen. They also showed the genes in both salicylic acid (SA) and jasmonic acid (JA) pathways were downregulated in *sal1*, suggesting the involvement of *SAL1* in plant immunity (Ishiga et al., 2017).

In GSL biosynthesis, several groups of GSLs differ from their corresponding precursors. Collectively, there are three GSL groups: aliphatic GSLs produced from methionine, alanine, leucine, isoleucine or valine; indolic GSLs produced from tryptophan; and benzyl GSLs produced from phenylalanine or tyrosine (Barba et al., 2016; Seo & Kim, 2017; Harun et al., 2020). However, the aliphatic GSL biosynthesis also needs another crucial step, which is the side-chain elongation in the chloroplast, followed by core structure synthesis in the cytoplasm (Sønderby, Geu-flores & Halkier, 2010; Borpatragohain et al., 2016). Based on Fig. 4, valine, leucine and isoleucine biosynthesis are the most enriched among the top six clusters generated in this study. Several GSL biosynthetic genes (*BCAT3*, *IPM11*, *IPM12*, and *IIL1*) are involved in the side-chain elongation process in the biosynthesis of valine, leucine and isoleucine. These genes are known as aliphatic GSL biosynthetic genes. The isopropylmalate dehydrogenases have been reported in side-chain GSL biosynthesis involving oxidative decarboxylation that produces a chain-elongated 2-oxo acid GSL. In this study, isopropylmalate dehydrogenase 3 (*IMDH3*) was also identified as a potential GSL gene which directly involved in the biosynthesis of valine, leucine and isoleucine. Previous T-DNA mutant studies on *Arabidopsis* *IMDH1*, *IMDH2*, and *IMDH3* showed a significantly decreased level of aliphatic GSLs and leucine in *IMDH1*, suggesting a clear role of *IMDH1* in catalyzing the oxidative decarboxylation step of aliphatic GSL biosynthesis (He et al., 2011b; Lee, Nwumeh & Jez, 2016) and double mutant of both *IMDH2* and *IMDH3* showed alteration in pollen and embryo sac growth, suggesting their correlation between leucine biosynthesis and the gametophyte formation in *Arabidopsis* (He et al., 2011a).

Tryptophan metabolism pathway was found in the gene network as shown in Fig. 4 where all genes directly linked to tryptophan metabolism are involved in indolic GSLs (Seo & Kim, 2017; Harun et al., 2020). The genes can be grouped based on their function in the indolic GSL biosynthesis: GSL core structure synthesis (CYP79B2, CYP79B3, CYP83B1, SUR1, and UGT74B1), as well as GSL degradation (TGG1, TGG2, and TGG4). CYP79B2 and CYP79B3 are P450 enzymes that catalyze the production of aldoximes from the tryptophan derivatives (Mikkelsen & Halkier, 2003). Next, another member of the P450 family, CYP83B1 catalyzes the oxidation of aldoximes into nitrile oxides, which is another crucial step in the GSL core structure synthesis (Naur et al., 2003). The myrosinase enzymes (TGG, EC 3.2.1.147) facilitates the production of active GSLs (bioactive isothiocyanates, nitriles, thiocyanates, and epithionitriles) in damaged plant cells during pest attacks or food preparation. Such activated GSL products have protective roles in plants against the biotic and abiotic stresses (Halkier & Gershenzon, 2006; Liu et al., 2020). The variations of the products are based on the GSLs side chain composition and the involvement of the myrosinase interacting proteins in the GSL-myrosinase system (Wittstock et al., 2016; Chhajed et al., 2019).

Lastly, the GSL biosynthesis linked with both aliphatic and indolic GSLs was also enriched in the gene network (Fig. 4). The aliphatic side-chain elongation genes were MAM1, MAM3, BCAT3, BCAT4, IPMI1, IPMI2, and IIL1. Methylthioalkylmalate synthase 1 (MAM1) is among the earliest aliphatic GSL genes identified in 2001. It is located in the GSL-ELONG locus, known to control the biosynthesis of GSL (Kroymann et al., 2001). Another gene that controls aliphatic side-chain elongation GSL is branched-chain aminotransferase 4 (BCAT4), and *in vivo* analysis of BCAT4 knockout plants showed significantly reduced Met-derived aliphatic GSL production (Schuster et al., 2006). Another group of enzyme is involved in the core structure synthesis (CYP79F1, CYP83A1, CYP79A2, SUR1, UGT74B1, SOT16, SOT17, and SOT18). In the aliphatic GSL core structure synthesis, CYP79F1 converts all chain-elongated methionine derivatives into aldoximes (Hansen et al., 2001). The aldoximes are then oxidized by CYP83A1 into activated aci-nitro compounds (Naur et al., 2003) CYP79A2 is specifically involved in benzyl GSL production as identified from the engineering of benzyl GSL pathway in *Nicotiana benthamiana* (Wittstock & Halkier, 2000; Geu-Flores et al., 2009). SUR1 and UGT74B1 are involved in aliphatic and indolic GSL biosynthesis *via* C-S lyase reaction and glycosylation, respectively. The end-product, desulfoGSLs would undergo sulfonation *via* the sulfotransferases (SOT16, SOT17, and SOT18), producing the GSL core structure in the cytosol (Piotrowski et al., 2004; Harun et al., 2020).

All enriched pathways identified in the gene network are known to be involved in GSL biosynthesis and metabolism. However, there are several known TFs and biosynthetic genes that are previously not linked with the terms in the pathway enrichment due to the usage of KEGG pathway in the ClueGO apps. We realized that side-chain modification was unavailable in the KEGG database even though side-chain modification pathway is one of the crucial step in GSL biosynthesis where the production of side chains would determine the biological functions of the activated GSL end products (Harun et al., 2020).

The flavin monooxygenases ( $FMO_{GS-OX1}$ ,  $FMO_{GS-OX2}$ ,  $FMO_{GS-OX3}$ , and  $FMO_{GS-OX5}$ ) are the side-chain modification enzymes that catalyze S-oxygenation process of methylthioalkyl GSL to methylsulfinylalkyl GSL. This process influences further modifications of GSL core structure that later produces the final GSL hydrolysis products in GSL biosynthesis (Hansen *et al.*, 2007; Kong *et al.*, 2016). Other missing components in the metabolic pathway database are the regulatory genes that encode for TFs and transporter proteins, both are crucial in the multi-component pathways like GSL biosynthesis. Thus, by referring to the latest articles and databases such as our in-house database, SuCCombase, the latest information of GSL components could improve the constructed GSL gene network.

The discovery of potential GSL genes (*MTO1*, *SIR*, *SAL1*, and *IMDH3*) from the pathway enrichment suggest their contribution in the aliphatic GSL biosynthesis based on their co-expression with known aliphatic GSL biosynthetic genes obtained from our proposed approach. Several potential genes unlinked to any enriched GSL pathway but might be involved in GSL biosynthesis: *MVP1*, *T19K24.17*, *MRSA2*, *ASP4*, *At1g21440*, *HMT3*, *At3g47420*, *PS1*, and *At3g14220* were identified. These genes were not mentioned in any known GSL pathway in the KEGG PATHWAY database; however, they were observed in the significantly enriched clusters obtained from the calculated AUC value. This finding is worth for a molecular validation *e.g.*, mutant studies, functional studies of downstream genes, and targeted metabolomics approach in order to prove their involvement in the GSL biosynthesis.

## CONCLUSIONS

Previously, we have successfully identified a novel GSL gene *ADAP* (Table S1) in the GSL biosynthesis *via* the similar approach and carried out an experimental validation of its involvement in the biosynthesis (Harun *et al.*, 2021). Thirteen potential GSL genes from the top six significant clusters: *IMDH3*, *MVP1*, *T19K24.17*, *MRSA2*, *SIR*, *ASP4*, *MTO1*, *At1g21440*, *HMT3*, *At3g47420*, *PS1*, *SAL1*, and *At3g14220* were identified from the GSL enriched clusters. Both *MRSA2* and *At1g21440* were the identified co-expressed genes in an aliphatic GSL co-expression network conducted in a previous study giving a high possibility of these genes as the potential GSL genes. Pathway enrichment analysis show direct involvement of four potential genes (*MTO1*, *SIR*, *SAL1*, and *IMDH3*) in the GSL biosynthesis-related pathways; sulfur metabolism and valine, leucine and isoleucine biosynthesis. This work demonstrated the application of network biology approach in the identification of missing genes and their related pathways. The combinatorial approach using graph clustering, Fisher's exact test, and ROC analysis on the constructed network biology can be used as an alternative technique to search for missing genes that cannot be found using the traditional sequence-based searching approach. This computational pipeline will benefit the scientific community in search for valuable information in the new gene discovery efforts. Furthermore, accurate knowledge on these genes is beneficial to plant scientists in the creation of genetic resources for crop improvement.

## ACKNOWLEDGEMENTS

We thank the Centre for Bioinformatics Research (CBR), Institute of Systems Biology (INBIOSIS), Universiti Kebangsaan Malaysia and Computational Systems Biology Lab, Nara Institute of Science and Technology (NAIST) for the computational facilities.

## ADDITIONAL INFORMATION AND DECLARATIONS

### Funding

This work was supported by the Malaysian Ministry of Higher Education (ERGS/1/2013/STG07/UKM/02/3) awarded to Zeti-Azura Mohamed-Hussein. Sarahani Harun is funded by JASSO for her attachment at NAIST to perform this experiment. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Grant Disclosures

The following grant information was disclosed by the authors:

Malaysian Ministry of Higher Education: ERGS/1/2013/STG07/UKM/02/3.  
JASSO.

### Competing Interests

The authors declare that they have no competing interests.

### Author Contributions

- Sarahani Harun conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Nor Afiqah-Aleng performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Mohammad Bozlul Karim performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Md Altaf Ul Amin conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Shigehiko Kanaya conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Zeti-Azura Mohamed-Hussein conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

### Data Availability

The following information was supplied regarding data availability:

The 127 potential GSL genes and the 148 significant clusters are available in the [Supplemental Files](#).

## Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.11876#supplemental-information>.

## REFERENCES

- Aarabi F, Kusajima M, Tohge T, Konishi T, Gigolashvili T, Takamune M, Sasazaki Y, Watanabe M, Nakashita H, Fernie AR, Saito K, Takahashi H, Hubberten H-M, Hoefgen R, Maruyama-Nakashita A. 2016. Sulfur deficiency-induced repressor proteins optimize glucosinolate biosynthesis in plants. *Science Advances* 2(10):e1601087 DOI 10.1126/sciadv.1601087.
- Afiqah-Aleng N, Altaf-UI-Amin M, Kanaya S, Mohamed-Hussein Z-A. 2020. Polycystic ovarian syndrome novel proteins and significant pathways identified using graph clustering approach. *Reproductive BioMedicine Online* 40(2):319–330 DOI 10.1016/j.rbmo.2019.11.012.
- Agerbirk N, Olsen CE. 2012. Glucosinolate structures in evolution. *Phytochemistry* 77(S1):16–45 DOI 10.1016/j.phytochem.2012.02.005.
- Altaf-UI-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S. 2006. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics* 7(1):207 DOI 10.1186/1471-2105-7-207.
- Altaf-UI-Amin M, Wada M, Kanaya S. 2012. Partitioning a PPI network into overlapping modules constrained by high-density and periphery tracking. *ISRN Biomathematics* 2012(6761):1–11 DOI 10.5402/2012/726429.
- Aoki Y, Okamura Y, Tadaka S, Kinoshita K, Obayashi T. 2016. ATTED-II in 2016: a plant coexpression database towards special online collection. *Plant & Cell Physiology* 57(1):1–9 DOI 10.1093/pcp/pcv165.
- Barba FJ, Nikmaram N, Roohinejad S, Khelifa A, Zhu Z, Koubaa M. 2016. Bioavailability of glucosinolates and their breakdown products: impact of processing. *Frontiers in Nutrition* 3:24 DOI 10.3389/fnut.2016.00024.
- Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. 2013. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* 41(D1):D991–D995 DOI 10.1093/nar/gks1193.
- Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Bye-Ajee H, Cowley A, Da Silva A, De Giorgi M, Dogan T, Fazzini F, Castro LG, Figueira L, Garmiri P, Georghiou G, Gonzalez D, Hatton-Ellis E, Li W, Liu W, Lopez R, Luo J, Lussi Y, MacDougall A, Nightingale A, Palka B, Pichler K, Poggioli D, Pundir S, Pureza L, Qi G, Rosanoff S, Saidi R, Sawford T, Shypitsyna A, Speretta E, Turner E, Tyagi N, Volynkin V, Wardell T, Warner K, Watkins X, Zaru R, Zellner H, Xenarios I, Bougueleret L, Bridge A, Poux S, Redaschi N, Aimo L, ArgoudPuy G, Auchincloss A, Axelsen K, Bansal P, Baratin D, Blatter MC, Boeckmann B, Bolleman J, Boutet E, Breuza L, Casal-Casas C, De Castro E, Coudert E, Cucho B, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gasteiger E, Gehant S, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Jungo F, Keller G, Lara V, Lemercier P, Lieberherr D, Lombardot T, Martin X, Masson P, Morgat A, Neto T, Noupik N, Paesano S, Pedruzzi I, Pilbout S, Pozzato M, Pruess M, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Wu CH, Arighi CN, Arminski L. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* 45(D1):D158–D169 DOI 10.1093/nar/gkw1099.



- Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman W, Pagès F, Trajanoski Z, Galon J, Team A, Immunology IC, Descartes UP. 2009.** ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25(8)**:1091–1093 DOI [10.1093/bioinformatics/btp101](https://doi.org/10.1093/bioinformatics/btp101).
- Bischoff KL. 2016.** Glucosinolates. In: *Nutraceuticals: Efficacy, Safety and Toxicity*. Amsterdam, Netherlands: Elsevier Inc, 551–554.
- Blažević I, Montaut S, Burčul F, Olsen CE, Burow M, Rollin P, Agerbirk N. 2019.** Glucosinolate structural diversity, identification, chemical synthesis and metabolism in plants. *Phytochemistry* **169(4)**:112100 DOI [10.1016/j.phytochem.2019.112100](https://doi.org/10.1016/j.phytochem.2019.112100).
- Borpatragohain P, Rose TJ, King GJ, King GJ. 2016.** Fire and brimstone: molecular interactions between sulfur and glucosinolate biosynthesis in model and crop brassicaceae. *Frontiers in Plant Science* **7(e27740)**:1–18 DOI [10.3389/fpls.2016.01735](https://doi.org/10.3389/fpls.2016.01735).
- Bozlul Karim M, Wakamatsu N, Altaf-Ul-Amin M. 2017.** DPclusOST: a software tool for general purpose graph clustering. *Journal of Computer Aided Chemistry* **18(0)**:76–93 DOI [10.2751/jcac.18.76](https://doi.org/10.2751/jcac.18.76).
- Burow M, Atwell S, Francisco M, Kerwin RE, Halkier BA, Kliebenstein DJ. 2015.** The glucosinolate biosynthetic gene AOP2 mediates feed-back regulation of jasmonic acid signaling in Arabidopsis. *Molecular Plant* **8(8)**:1201–12012 DOI [10.1016/j.molp.2015.03.001](https://doi.org/10.1016/j.molp.2015.03.001).
- Chhajed S, Misra BB, Tello N, Chen S. 2019.** Chemodiversity of the glucosinolate-myrosinase system at the single cell type resolution. *Frontiers in Plant Science* **10**:120 DOI [10.3389/fpls.2019.00618](https://doi.org/10.3389/fpls.2019.00618).
- Davis J, Goadrich M. 2006.** The relationship between precision-recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning*, New York 233–240.
- de Anda-Jáuregui G. 2019.** Guideline for comparing functional enrichment of biological network modular structures. *Applied Network Science* **4(1)**:1–17 DOI [10.1007/s41109-019-0128-1](https://doi.org/10.1007/s41109-019-0128-1).
- Eguchi R, Karim MB, Hu P, Sato T, Ono N, Kanaya S, Altaf-Ul-Amin M. 2018.** An integrative network-based approach to identify novel disease genes and pathways: A case study in the context of inflammatory bowel disease. *BMC Bioinformatics* **19(1)**:1–12 DOI [10.1186/s12859-018-2251-x](https://doi.org/10.1186/s12859-018-2251-x).
- Falk KL, Tokuhisa JG, Gershenzon J. 2007.** The effect of sulfur nutrition on plant glucosinolate content: physiology and molecular mechanisms. *Plant Biology* **9(5)**:573–581 DOI [10.1055/s-2007-965431](https://doi.org/10.1055/s-2007-965431).
- Fisher RA. 1992.** Statistical methods for research workers. In: Samuel K, Johnson Norman L, eds. *Breakthroughs in Statistics*. Vol. II. Berlin, Germany: Springer, 66–70.
- Fisher RA. 1922.** On the interpretation of  $\chi^2$  from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* **85(1)**:87–94 DOI [10.2307/2340521](https://doi.org/10.2307/2340521).
- Frerigmann H, Piślewska-Bednarek M, Sánchez-Vallet A, Molina A, Glawischnig E, Gigolashvili T, Bednarek P. 2016.** Regulation of pathogen-triggered tryptophan metabolism in *Arabidopsis thaliana* by MYB transcription factors and indole glucosinolate conversion products. *Molecular Plant* **9(5)**:682–695 DOI [10.1016/j.molp.2016.01.006](https://doi.org/10.1016/j.molp.2016.01.006).
- Gachon CMM, Langlois-Meurinne M, Henry Y, Saindrenan P. 2005.** Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: functional and evolutionary implications. *Plant Molecular Biology* **58(2)**:229–245 DOI [10.1007/s11103-005-5346-5](https://doi.org/10.1007/s11103-005-5346-5).
- Geu-Flores F, Nielsen MT, Nafisi M, Møldrup ME, Olsen CE, Motawia MS, Halkier BA. 2009.** Glucosinolate engineering identifies a  $\gamma$ -glutamyl peptidase. *Nature Chemical Biology* **5(8)**:575–577 DOI [10.1038/nchembio.185](https://doi.org/10.1038/nchembio.185).

- Halkier BA, Gershenzon J. 2006.** Biology and biochemistry of glucosinolates. *Annual Review of Plant Biology* 57(1):303–333 DOI 10.1146/annurev.arplant.57.032905.105228.
- Hansen BG, Kliebenstein DJ, Halkier BA, Ave OS. 2007.** Identification of a flavin-monooxygenase as the S-oxygenating enzyme in aliphatic glucosinolate biosynthesis in Arabidopsis. *The Plant Journal* 50(5):902–910 DOI 10.1111/j.1365-3113X.2007.03101.x.
- Hansen CH, Wittstock U, Olsen CE, Hick AJ, Pickett JA, Halkier BA. 2001.** Cytochrome P450 CYP79F1 from Arabidopsis catalyzes the conversion of dihomomethionine and trihomomethionine to the corresponding aldoximes in the biosynthesis of aliphatic glucosinolates. *The Journal of Biological Chemistry* 276(14):11078–11085 DOI 10.1074/jbc.M010123200.
- Harun S, Abdullah-Zawawi M-R, A-Rahman MRA, Muhammad NAN, Mohamed-Hussein Z-A. 2019.** SuCComBase: a manually curated repository of plant sulfur-containing compounds. *Database* baz021:10 DOI 10.1093/database/baz021.
- Harun S, Abdullah-Zawawi MR, Goh HH, Mohamed-Hussein ZA. 2020.** A comprehensive gene inventory for glucosinolate biosynthetic pathway in Arabidopsis thaliana. *Journal of Agricultural and Food Chemistry* 68(28):7281–7297 DOI 10.1021/acs.jafc.0c01916.
- Harun S, Rohani ER, Ohme-Takagi M, Goh H-H, Mohamed-Hussein Z-A. 2021.** ADAP is a possible negative regulator of glucosinolate biosynthesis in Arabidopsis thaliana based on clustering and gene expression analyses. *Journal of Plant Research* 134(2):327–339 DOI 10.1007/s10265-021-01257-9.
- He Y, Chen L, Zhou Y, Mawhinney TP, Chen B, Kang BH, Hauser BA, Chen S. 2011a.** Functional characterization of Arabidopsis thaliana isopropylmalate dehydrogenases reveals their important roles in gametophyte development. *New Phytologist* 189(1):160–175 DOI 10.1111/j.1469-8137.2010.03460.x.
- He Y, Galant A, Pang Q, Strul JM, Balogun SF, Jez JM, Chen S. 2011b.** Structural and functional evolution of isopropylmalate dehydrogenases in the leucine and glucosinolate pathways of Arabidopsis thaliana. *The Journal of Biological Chemistry* 286(33):28794–28801 DOI 10.1074/jbc.M111.262519.
- Herr I, Büchler MW. 2010.** Dietary constituents of broccoli and other cruciferous vegetables: Implications for prevention and therapy of cancer. *Cancer Treatment Reviews* 36(5):377–383 DOI 10.1016/j.ctrv.2010.01.002.
- Hirai MY, Fujiwara T, Awazuhara M, Kimura T, Noji M, Saito K. 2003.** Global expression profiling of sulfur-starved Arabidopsis by DNA microarray reveals the role of O-acetyl-L-serine as a general regulator of gene expression in response to sulfur nutrition. *Plant Journal* 33(4):651–663 DOI 10.1046/j.1365-3113X.2003.01658.x.
- Hirai MY, Klein M, Fujikawa Y, Yano M, Goodenowe DB, Yamazaki Y, Kanaya S, Nakamura Y, Kitayama M, Suzuki H, Sakurai N, Shibata D, Tokuhisa J, Reichelt M, Gershenzon J, Papenbrock J, Saito K. 2005.** Elucidation of gene-to-gene and metabolite-to-gene networks in Arabidopsis by integration of metabolomics and transcriptomics. *Journal of Biological Chemistry* 280(27):25590–25595 DOI 10.1074/jbc.M502332200.
- Hirai MY, Sugiyama K, Sawada Y, Tohge T, Obayashi T, Suzuki A, Araki R, Sakurai N, Suzuki H, Aoki K, Goda H, Nishizawa OI. 2007.** Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. *Proceedings of the National Academy of Sciences of The United States of America* 104(15):6478–6483 DOI 10.1073/pnas.0611629104.
- Ishiga Y, Watanabe M, Ishiga T, Tohge T, Matsuura T, Ikeda Y, Hoefgen R, Fernie AR, Mysore KS. 2017.** The SAL-PAP chloroplast retrograde pathway contributes to plant immunity

- by regulating glucosinolate pathway and phytohormone signaling. *Molecular Plant-Microbe Interactions* **30(10)**:829–841 DOI [10.1094/MPMI-03-17-0055-R](https://doi.org/10.1094/MPMI-03-17-0055-R).
- Jin J, Tian F, Yang D, Meng Y, Kong L, Luo J, Gao G. 2016. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research* **gkw982(D1)**:D1040–D1045 DOI [10.1093/nar/gkw982](https://doi.org/10.1093/nar/gkw982).
- Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45(D1)**:353–361 DOI [10.1093/nar/gkw1092](https://doi.org/10.1093/nar/gkw1092).
- Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44(D1)**:D457–D462 DOI [10.1093/nar/gkv1070](https://doi.org/10.1093/nar/gkv1070).
- Karim MB, Huang M, Ono N, Kanaya S, Amin MAU. 2020. BiClusO: a novel biclustering approach and its application to species-VOC relational data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **17(6)**:1955–1965 DOI [10.1109/TCBB.2019.2914901](https://doi.org/10.1109/TCBB.2019.2914901).
- Kliebenstein DJ, Kroymann J, Brown P, Figuth A, Pedersen D, Gershenzon J, Mitchell-olds T, Genetics D, Evolution DJK. 2001. Genetic control of natural variation in Arabidopsis glucosinolate accumulation. *Plant Physiology* **126(2)**:811–825 DOI [10.1104/pp.126.2.811](https://doi.org/10.1104/pp.126.2.811).
- Knill T, Schuster J, Reichelt M, Gershenzon J, Binder S. 2008. Arabidopsis branched-chain aminotransferase 3 functions in both amino acid and glucosinolate biosynthesis. *Plant Physiology* **146(3)**:1028–1039 DOI [10.1104/pp.107.111609](https://doi.org/10.1104/pp.107.111609).
- Kong W, Li J, Yu Q, Cang W, Xu R, Wang Y, Ji W. 2016. Two novel flavin-containing monooxygenases involved in biosynthesis of aliphatic glucosinolates. *Frontiers in Plant Science* **7(e2068)**:1–9 DOI [10.3389/fpls.2016.01292](https://doi.org/10.3389/fpls.2016.01292).
- Koprivova A, Kopriva S. 2014. Molecular mechanisms of regulation of sulfate assimilation: first steps on a long road. *Frontiers in Plant Science* **5(e1001193)**:1–11 DOI [10.3389/fpls.2014.00589](https://doi.org/10.3389/fpls.2014.00589).
- Kroymann J, Textor S, Tokuhisa JG, Falk KL, Bartram S, Gershenzon J, Mitchell-olds T. 2001. A gene controlling variation in Arabidopsis glucosinolate composition is part of the methionine chain elongation pathway. *Plant Physiology* **127**:1077–1088 DOI [10.1104/pp.010416.1](https://doi.org/10.1104/pp.010416.1).
- Lai K-C, Huang A-C, Hsu S-C, Kuo C-L, Yang J-S, Wu S-H, Chung J-G. 2010. Benzyl isothiocyanate (BITC) inhibits migration and invasion of human colon cancer HT29 cells by inhibiting matrix metalloproteinase-2/-9 and Urokinase Plasminogen (uPA) through PKC and MAPK signaling pathway. *Journal of Agricultural and Food Chemistry* **58(5)**:2935–2942 DOI [10.1021/jf9036694](https://doi.org/10.1021/jf9036694).
- Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E. 2012. The Arabidopsis information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* **40(D1)**:D1202–D1210 DOI [10.1093/nar/gkr1090](https://doi.org/10.1093/nar/gkr1090).
- Lee SG, Nwumeh R, Jez JM. 2016. Structure and mechanism of isopropylmalate dehydrogenase from Arabidopsis thaliana: insights on leucine and aliphatic glucosinolate biosynthesis. *Journal of Biological Chemistry* **291(26)**:13421–13430 DOI [10.1074/jbc.M116.730358](https://doi.org/10.1074/jbc.M116.730358).
- Lee T, Yang S, Kim E, Ko Y, Hwang S, Shin J, Shim E, Shim H, Kim H, Kim C, Lee I. 2014. AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. *Nucleic Acids Research* **43(D1)**:1–7 DOI [10.1093/nar/gku1053](https://doi.org/10.1093/nar/gku1053).

- Liu Y, Rossi M, Liang X, Zhang H, Zou L, Ong CN. 2020. An integrated metabolomics study of glucosinolate metabolism in different brassicaceae genera. *Metabolites* **10**(8):1–17 DOI [10.3390/metabo10080313](https://doi.org/10.3390/metabo10080313).
- Megna BW, Carney PR, Nukaya M, Geiger P, Kennedy GD. 2016. Indole-3-carbinol induces tumor cell death: function follows form. *Journal of Surgical Research* **204**(1):47–54 DOI [10.1016/J.JSS.2016.04.021](https://doi.org/10.1016/J.JSS.2016.04.021).
- Metz CE. 1978. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8**(4):283–298 DOI [10.1016/S0001-2998\(78\)80014-2](https://doi.org/10.1016/S0001-2998(78)80014-2).
- Miao H, Cai C, Wei J, Huang J, Chang J, Qian H, Zhang X, Zhao Y, Sun B, Wang B, Wang Q. 2016. Glucose enhances indolic glucosinolate biosynthesis without reducing primary sulfur assimilation. *Scientific Reports* **6**(1):1–12 DOI [10.1038/srep31854](https://doi.org/10.1038/srep31854).
- Mikkelsen MD, Halkier BA. 2003. Metabolic engineering of valine-and isoleucine-derived glucosinolates in Arabidopsis expressing CYP79D2 from cassava. *Plant Physiology* **131**(2):773–779 DOI [10.1104/pp.013425](https://doi.org/10.1104/pp.013425).
- Montejo J, Zuberi K, Rodriguez H, Bader GD, Morris Q. 2014. GeneMANIA: fast gene network construction and function prediction for Cytoscape. *F1000Research* **153**:1–7 DOI [10.12688/f1000research](https://doi.org/10.12688/f1000research).
- Morikawa-Ichinose T, Kim SJ, Allahham A, Kawaguchi R, Maruyama-Nakashita A. 2019. Glucosinolate distribution in the aerial parts of sel1-10, a disruption mutant of the sulfate transporter SULTR1; 2, in mature arabidopsis thaliana plants. *Plants* **88**:95 DOI [10.3390/plants8040095](https://doi.org/10.3390/plants8040095).
- Mueller LA, Zhang P, Rhee SY. 2003. AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiology* **132**(2):453–460 DOI [10.1104/pp.102.017236](https://doi.org/10.1104/pp.102.017236).
- Naur P, Petersen BL, Mikkelsen MD, Bak S, Rasmussen H, Olsen CE, Halkier BA. 2003. CYP83A1 and CYP83B1, two nonredundant cytochrome P450 enzymes metabolizing oximes in the biosynthesis of glucosinolates in Arabidopsis. *Plant Physiology* **133**:63–72 DOI [10.1104/pp.102.019240.1](https://doi.org/10.1104/pp.102.019240.1).
- Nikiforova V, Freitag J, Kempa S, Adamik M, Hesse H, Hoefgen R. 2003. Transcriptome analysis of sulfur depletion in Arabidopsis thaliana: interlacing of biosynthetic pathways provides response specificity. *Plant Journal* **33**(4):633–650 DOI [10.1046/j.1365-3113X.2003.01657.x](https://doi.org/10.1046/j.1365-3113X.2003.01657.x).
- Pathak RR, Jangam AP, Malik A, Sharma N, Jaiswal DK, Raghuram N. 2020. Transcriptomic and network analyses reveal distinct nitrate responses in light and dark in rice leaves (*Oryza sativa* Indica var. Panvel1). *Scientific Reports* **10**(1):12228 DOI [10.1038/s41598-020-68917-z](https://doi.org/10.1038/s41598-020-68917-z).
- Piotrowski M, Schemenewitz A, Lopukhina A, Mu A, Janowitz T, Weiler EW, Oecking C. 2004. Desulfoglucosinolate sulfotransferases from *Arabidopsis thaliana* catalyze the final step in the biosynthesis of the glucosinolate core structure. *The Journal of Biological Chemistry* **279**(49):50717–50725 DOI [10.1074/jbc.M407681200](https://doi.org/10.1074/jbc.M407681200).
- Piślewska-Bednarek M, Singkaravanit-Ogawa S, Schulze-Lefert P, Sanchez-Vallet A, Nakano RT, Bednarek P, Hiruma K, Pastorczyk M, Molina A, Takano Y, Ciesiolka D. 2017. Glutathione transferase U13 functions in pathogen-triggered glucosinolate metabolism. *Plant Physiology* **176**(1):538–551 DOI [10.1104/pp.17.01455](https://doi.org/10.1104/pp.17.01455).
- Quintero FJ, Garcíadeblás B, Rodríguez-Navarro A. 1996. The SAL1 gene of Arabidopsis, encoding an enzyme with 3'(2'),5'-bisphosphate nucleotidase and inositol polyphosphate 1-phosphatase activities, increases salt tolerance in yeast. *Plant Cell* **8**(3):529–537 DOI [10.1105/tpc.8.3.529](https://doi.org/10.1105/tpc.8.3.529).
- Redovniković IR, Glivetic T, Delonga K, Jasna V-F. 2008. Glucosinolates and their potential role in plant. *Periodicum Biologorum* **110**:297–309.

- Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J, Keays M, Kurbatova N, Malone J, Mani R, Mupo A, Pereira RP, Pilicheva E, Rung J, Sharma A, Tang YA, Ternent T, Tikhonov A, Welter D, Williams E, Brazma A, Parkinson H, Sarkans U. 2013. ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Research* 41(D1):D987–D990 DOI 10.1093/nar/gks1174.
- Sawada Y, Kuwahara A, Nagano M, Narisawa T, Sakata A, Saito K, Yokota Hirai M. 2009a. Omics-based approaches to methionine side chain elongation in Arabidopsis: characterization of the genes encoding methylthioalkylmalate isomerase and methylthioalkylmalate dehydrogenase. *Plant & Cell Physiology* 50(7):1181–1190 DOI 10.1093/pcp/pcp079.
- Sawada Y, Toyooka K, Kuwahara A, Sakata A, Nagano M, Saito K, Hirai MY. 2009b. Arabidopsis bile acid: sodium symporter family protein 5 is involved in methionine-derived glucosinolate biosynthesis. *Plant & Cell Physiology* 50(9):1579–1586 DOI 10.1093/pcp/pcp110.
- Schuster J, Knill T, Reichelt M, Gershenzon J, Binder S. 2006. BRANCHED-CHAIN AMINOTRANSFERASE4 is part of the chain elongation pathway in the biosynthesis of methionine-derived glucosinolates in Arabidopsis. *The Plant Cell* 18(10):2664–2679 DOI 10.1105/tpc.105.039339.
- Seo MS, Kim JS. 2017. Understanding of MYB transcription factors involved in glucosinolate biosynthesis in Brassicaceae. *Molecules* 22(9):1549 DOI 10.3390/molecules22091549.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 13(11):2498–2504 DOI 10.1101/gr.1239303.
- Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCr: visualizing classifier performance in R. *Bioinformatics* 21(20):3940–3941.
- Sønderby IE, Geu-flores F, Halkier BA. 2010. Biosynthesis of glucosinolates—gene discovery and beyond. *Trends in Plant Science* 15(5):283–290 DOI 10.1016/j.tplants.2010.02.005.
- Sønderby IE, Geu-Flores F, Halkier BA. 2010. Biosynthesis of glucosinolates—gene discovery and beyond. *Trends in Plant Science* 15(5):283–290 DOI 10.1016/j.tplants.2010.02.005.
- Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, Von Mering C. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* 43(D1):D447–D452 DOI 10.1093/nar/gku1003.
- Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Von Mering C. 2019. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* 47(D1):D607–D613 DOI 10.1093/nar/gky1131.
- Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research* 45(D1):D362–D368 DOI 10.1093/nar/gkw937.
- Taiz L, Zeiger E. 2010. *Plant physiology*. Sunderland, MA: Sinauer Associates, Inc.
- Takahashi H, Kopriva S, Giordano M, Saito K, Rüdiger H. 2011. Sulfur assimilation in photosynthetic organisms: molecular functions and regulations of transporters and assimilatory enzymes. *Annual Review of Plant Biology* 62(1):157–186 DOI 10.1146/annurev-arplant-042110-103921.



- Tang F, Xie C, Huang D, Wu Y, Zeng M, Yi L, Wang Y, Mei W, Cao Y, Sun L. 2011.** Novel potential markers of nasopharyngeal carcinoma for diagnosis and therapy. *Clinical Biochemistry* **44**(8–9):711–718 DOI [10.1016/j.clinbiochem.2011.03.025](https://doi.org/10.1016/j.clinbiochem.2011.03.025).
- Tang L, Zirpoli GR, Guru K, Moysich KB, Zhang Y, Ambrosone CB, McCann SE. 2010.** Intake of cruciferous vegetables modifies bladder cancer survival. *Cancer Epidemiology Biomarkers and Prevention* **19**(7):1806–1811 DOI [10.1158/1055-9965.EPI-10-0008](https://doi.org/10.1158/1055-9965.EPI-10-0008).
- Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q. 2010.** The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research* **38**(suppl\_2):W214–W220 DOI [10.1093/nar/gkq537](https://doi.org/10.1093/nar/gkq537).
- Wink M. 2015.** Modes of action of herbal medicines and plant secondary metabolites. *Medicines* **2**(3):251–286 DOI [10.3390/medicines2030251](https://doi.org/10.3390/medicines2030251).
- Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A. 2017.** A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. *The Plant Cell* **29**(5):944–959 DOI [10.1105/tpc.17.00009](https://doi.org/10.1105/tpc.17.00009).
- Wittstock U, Halkier BA. 2000.** Cytochrome P450 CYP79A2 from *Arabidopsis thaliana* L. catalyzes the conversion of L-phenylalanine to phenylacetaldoxime in the biosynthesis of benzylglucosinolate. *The Journal of Biological Chemistry* **275**:14659–14666.
- Wittstock U, Meier K, Dörr F, Ravindran BM. 2016.** NSP-dependent simple nitrile formation dominates upon breakdown of major aliphatic glucosinolates in roots, seeds, and seedlings of *Arabidopsis thaliana* Columbia-0. *Frontiers* **7**:1821 DOI [10.3389/fpls.2016.01821](https://doi.org/10.3389/fpls.2016.01821).
- Yatusevich R, Mugford SG, Matthewman C, Gigolashvili T, Frerigmann H, Delaney S, Koprivova A, Flu U. 2010a.** Genes of primary sulfate assimilation are part of the glucosinolate biosynthetic network in *Arabidopsis thaliana*. *The Plant Journal* **62**(1):1–11 DOI [10.1111/j.1365-313X.2009.04118.x](https://doi.org/10.1111/j.1365-313X.2009.04118.x).
- Yatusevich R, Mugford SG, Matthewman C, Gigolashvili T, Frerigmann H, Delaney S, Koprivova A, Flu U. 2010b.** Genes of primary sulfate assimilation are part of the glucosinolate biosynthetic network in *Arabidopsis thaliana*. *The Plant Journal* **62**(1):1–11 DOI [10.1111/j.1365-313X.2009.04118.x](https://doi.org/10.1111/j.1365-313X.2009.04118.x).