Chapter

Computational Analysis of Rice Transcriptomic and Genomic Datasets in Search for SNPs Involved in Flavonoid Biosynthesis

Rabiatul-Adawiah Zainal-Abidin
and Zeti-Azura Mohamed-Hussein

Abstract

This chapter describes the computational approach used in analyzing rice transcriptomics and genomics data to identify and annotate potential single nucleotide polymorphism (SNPs) as potential biomarker in the production of flavonoid. SNPs play a role in the accumulation of nutritional components (e.g. antioxidants), and flavonoid is one of them. However, the number of identified SNPs associated with flavonoid nutritional trait is still limited. We develop a knowledge-based bioinformatic workflow to search for specific SNPs and integration analysis on the SNPs and their co-expressed genes to investigate their influence on the gain/loss of functional genes that are involved in the production of flavonoids. Raw files obtained from the functional genomics studies can be analyzed in details to obtain a useful biological insight. Different tools, algorithms and databases are available to analyze the ontology, metabolic and pathway at the molecular level in order to observe the effects of gene and protein expression. The usage of different tools, algorithms and databases allows the integration, interpretation and the inference of analysis to provide better understanding of the biological meaning of the resutls. This chapter illustrates how to select and bring together several software to develop a specific bioinformatic workflow that processes and analyses omics data. The implementation of this bioinformatic workflow revealed the identification of potential flavonoid biosynthetic genes that can be used as guided-gene to screen the single nucleotide polymorphisms (SNPs) in the flavonoid biosynthetic genes from genome and transcriptomics data.

Keywords: SNPs, comparative genomics, transcriptome, nutritional traits, colored rice, integrative analysis, bioinformatics, flavonoid, genomics, single nucleotide polymorphism, transcriptomics, bioinformatics workflow

1. Introduction

In recent years, high-throughput omics technologies (i.e. genomics, proteomics, transcriptomics and/or metabolomics) provide unprecedented opportunities to discover potential genes, proteins, metabolites, pathways and molecular markers for

various applications. The availability of omics data produced from high-throughput omics technologies has facilitated the molecular and genetic improvement of rice varieties with higher yield, quality, nutrient dan resistance to the biotic and abiotic stresses. However, past 15 years have seen the significant increase of omics data in volume and types. This event has challenged the researchers to extract and decipher invaluable information encoded in the data. This challenge can be addressed with the use of bioinformatics in analyzing, integrating and interpreting these massive omics data. There are various bioinformatic tools and algorithms that can be used by the researchers to process, interpret and integrate data in a more efficient and reproducible way. However, lack of connectivity between various tools and algorithms complicates the process of extracting and deciphering the data. Hence there is the need to find procedures to connect these tools to develop workflows that can connect and bring together different techniques that are able to exploit data at many levels.

Here we describe a computational workflow composed of different bioinformatic tools that exploits data from large-scale gene expression experiments and contextualize them at many biological levels. To illustrate the relevance of our workflow, we applied it to data from rice varieties datasets in search for potential SNPs that are associated with flavonoid biosynthetic genes. The workflow started with identification of known flavonoid biosynthetic genes from published articles and database search using several genome and pathway databases such as. KEGG, PlantReactome, RiceCyc) and similarity search analysis.

The potential flavonoid biosynthetic genes were used as a guide-gene to screen for single nucleotide polymorphisms (SNPs) in the flavonoid biosynthetic genes from the genomics and transcriptomics data. Integration of SNP and co-expressed genes was performed via network analysis. The transcriptomics data was used to construct the gene co-expression network followed by the mapping of SNPs onto it. A pathway-network analysis was performed to interpret the biological information related to the flavonoid pathway-network. All information generated from these computational analyses are stored in *MyNutRice*Base (http://www.mynutricebase.org) for knowledge sharing and future use in the functional genomics study and the development of molecular markers.

2. Overview of flavonoids in omics and rice breeding improvement

Several types of rice breeding traits such as disease resistance, drought tolerance, salinity tolerance, grain quality, high nutritional content are currently pursue by the breeders and farmers in their effort to improve the traits of rice varieties. However, nowadays breeding for improved rice varieties with high nutritional content has attracted interest among breeders, geneticists and nutritionists. High protein [1], carotenoid [2], micronutrients [3] and antioxidant content [4] are among the preferred nutritional contents improved in rice varieties. Rice breeding for high nutritional content has been carried out extensively by bio-fortification [5, 6] and/or genetic engineering [7]. The improvement of nutritional content in rice is essential due to their benefits for human health. Lack of specific nutrition can lead to several diseases and malnutrition.

Flavonoids are secondary metabolites commonly produced in flowers, fruits, vegetables and pigmented rice. Flavonoids are known as a potent antioxidant beneficial for human health. Consumption of foods with high antioxidant contents may lower the risk of cardiovascular disease, type II diabetes and colon cancer [8]. Additionally, flavonoid has shown to improve plant resistance against abiotic and biotic stress [9].

The flavonoid biosynthetic pathways are found in several crops such as maize, tomato and rice [10]. Genes encoding enzymes involved in the flavonoid biosynthetic pathways are categorized into general phenylpropanoid, early biosynthetic genes (EBGs), late biosynthetic genes (LBGs) and transcription factors [11]. General phenylpropanoid contains three major genes such as phenylalanine ammonia-lyase (PAL), cinnamic acid 4-hydroxylase (C4H) and 4-coumarate CoA ligase (4CL). The EBGs include chalcone synthase (CHS), chalcone isomerase (CHI), flavanone 3-hydroxylase (F3H) and flavanone 3′-hydroxylase (F3'H) [11]. Genes in the general phenylpropanoid category and EBGs are upstream genes that initiate the flavonoid biosynthetic pathways and responsible in the production of secondary metabolites [11]. The LBGs such as dihydroflavonol reductase (DFR), leucoanthocyanidin reductase (LAR), UDP-glucose flavonoid 3-O-glucosyl transferase (UGT) and leucoanthocyanidin oxidase (LDOX) lead to the production of anthocyanin [11]. Genes categorized in EBGs and LBGs are also recognized as structural genes [11].

Four transcription factors involved in the flavonoid biosynthetic pathways are *Kala4* (Os04g0557500), *Rc* (Os07g0211500), *R2R3-MYB* (Os06g0205100) and *WD40* (Os02g0682500). *Kala4* encodes a basic-helix–loop–helix (bHLH) and it activates late biosynthetic genes to produce black pigmentation or anthocyanin in the seed pericarps [12]. *Rc* gene encodes for bHLH that regulates the *Rd* (Os01g0633500) expression to produce red pigmentation [13]. However, *Rc* expression without the participation of *Rd* has resulted to brown pigmentation [13]. Besides the existence of *R2R3-MYB* and *WD40* in the flavonoid biosynthesis, they also responsible to regulate the pigmentation in purple leaves [14, 15].

Anthocyanin and proanthocyanidins are two major flavonoid compounds [11]. Pigmented rice (black and red) is enriched with antioxidant due to the presence of anthocyanin and proanthocyanidin [16]. Understanding the genetic basis of pigmented rice varieties is essential to develop rice varieties with high antioxidant (flavonoid, anthocyanin, proanthocyanidin) content. A study by [17] has dissected the regulation of flavonoid biosynthesis in edible rice tissue using a metabolic engineering approach. Bioinformatics approach was used to identify key flavonoid structural genes in the rice genome by species comparison against maize and sorghum as sequence homologs [17]. A total of six genes encoding enzymes (CHS, CHI, F3H, F3'H, DFR and ANS) in the flavonoid biosynthetic pathway were selected for tBLASTN analysis against Nipponbare rice genome sequence. At least 66% amino acid identities were found in the rice genome sequences. The expression patterns of six flavonoid genes were analyzed to investigate the accumulation of flavonoids level in rice seedlings.

Previous study showed the correlation between sequence polymorphism and metabolite profiling affects the flavonoid accumumation between *indica* and *japonica* rice sub-species [18]. Different accumulation of flavonoids in these two rice sub-species might due to the variation in flavonoid biosynthetic genes [19, 20]. A quantitative trait loci analysis was performed to develop molecular markers associated with antioxidant content in rice [21]. The potential molecular markers associated with antioxidant content can be applied in the marker-assisted breeding towards the improvement of high-level antioxidant content in rice variety.

Understanding gene-based-SNP underlying flavonoid biosynthesis process is crucial in developing rice cultivars with higher flavonoid contents. Integration of single nucleotide polymorphisms (SNPs) and co-expressed genes can be used to identify causal SNPs and to prioritize the functional SNPs involved with flavonoid biosynthetic genes.

However, the study on molecular and genetic improvement of antioxidant content in rice via integration of multi-omics data is still limited. Current data is still

inadequate to be applied in rice breeding selection. Furthermore, this constraint limits detailed insight into the underlying mechanism and regulation of antioxidant content on the system level. The association of SNPs in the causative biosynthetic genes might be useful to uncover key alleles that influence the accumulation of flavonoid. Linking the SNPs with their co-expressed genes that involved in the flavonoid biosynthesis process could be a promising approach to prioritize the functional SNPs and causal genes to be used in the experimental validation towards the molecular and genetic improvement of rice variety enriched with flavonoid content.

3. Data analysis workflow

3.1 Data mining of potential genes and transcription factor related to flavonoid biosynthesis

Pubmed (https://pubmed.ncbi.nlm.nih.gov/), Web of Science (https://www.webofknowledge.com/) and Scopus (https://www.scopus.com/) were used systematically to search for publications related to flavonoid biosynthetic genes in *Oryza sativa* using the following keywords and their combinations such as "flavonoid", "rice", "SNP", "anthocyanin", "proanthocyanidin" and "pigmented". The keywords such as "flavonoid", "C4H", "4CL", "CHS", "F3H", "F3'H", "UGT", "DFR", "CHI", "PAL", "LAR", "ANS", "ANR", "LDOX", anthocyanin" and "proanthocyanidin" were used in the genome (i.e. RAPDB) and pathway (i.e. KEGG, RiceCyc, PlantReactome) databases to search for gene identifier, gene sequences and gene descriptions.

All genes that are related to the flavonoid genes were identified using similarity search analysis from OmicsBox (https://www.biobam.com/omicsbox/). BLASTX [22] analysis was performed with the e-value cut off of less than 1e-10 and sequence identity more than 75%. eggNOG [23] analysis was carried out to characterize whether the sequences are paralogous or orthologous.

Bibliomic, database and similarity search are the first steps in identifying candidate genes related to the biosynthesis pathways. These genes that can be used in downstream analysis1 such as evolutionary study [24] and identification of SNPs in the biosynthetic genes [25]. From these steps, a total of 95 flavonoid related genes (FRGs) and two transcription factors were successfully identified. Structural genes (i.e. F3H, CHI, CHS,DFR, LDOX) are found to be more conserved than transcription factors. These flavonoid related genes were used to screen the SNPs that reside in the sequences. **Figure 1** shows the three steps analysis workflow to identify the potential flavonoid related genes.

3.2 Mining of SNPs from genome and transcriptome data

Single nucleotide polymorphism (SNP) has been widely used as a genetic marker tool in crop improvement. Previous studies used SNPs to investigate the evolutionary relationship [26], to facilitate cultivar identification [27] and to associate SNPs with agronomic traits [28]. SNPs located in the intergenic and genic region [29]. The genic SNPs can be classified into the coding region, untranslated region (UTR) and intron. SNPs in the coding region are divided into synonymous and non-synonymous. The non-synonymous SNPs has two effects such as deleterious and tolerance that might represent causal genetic variation which could lead to the phenotypic consequences [30].

SNPs in the coding regions are usually associated with functional SNP and they can influence the phenotypic expression of agronomic traits such as

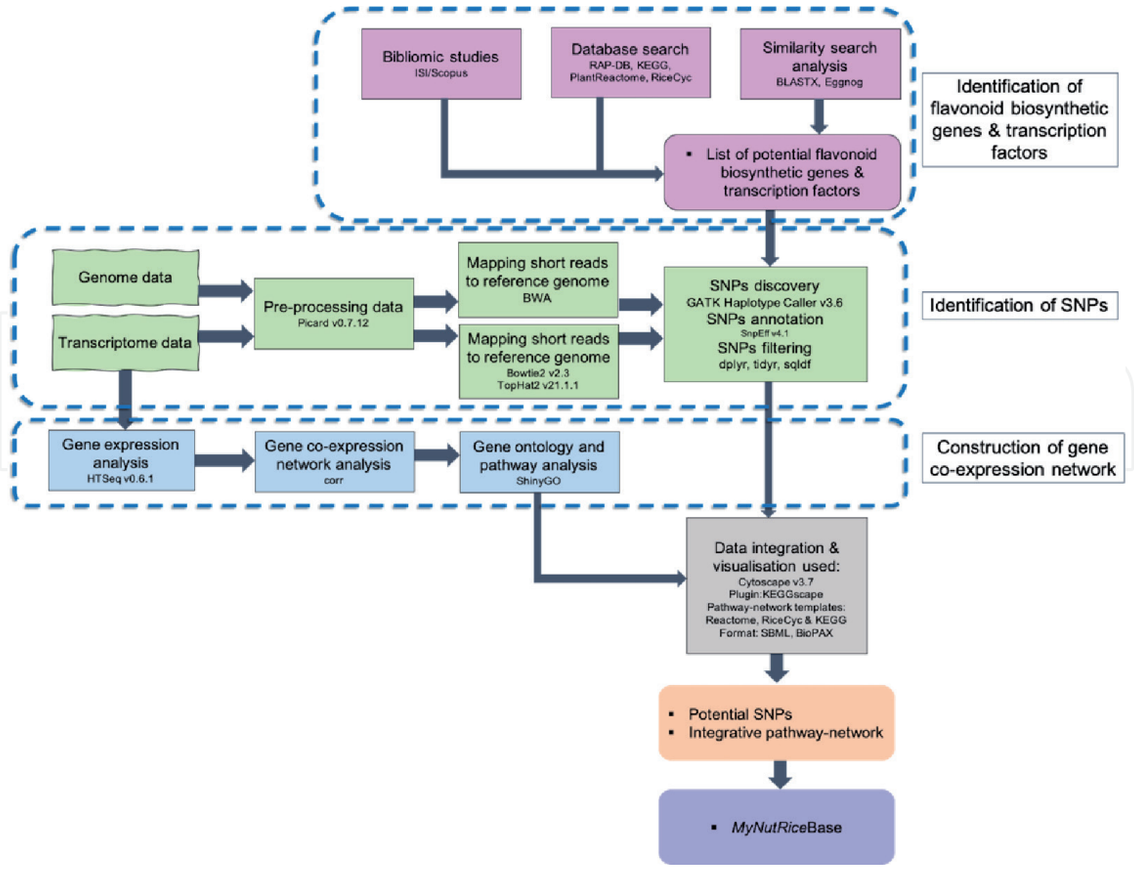

Figure 1.
Bioinformatic analysis workflow on the genome and transcriptomic datasets of six Malaysian rice varieties.

resistance to disease, resistance to drought and high nutritional content [31]. The non-synonymous SNPs might affect the protein function through amino acid substitution [32] whilst synonymous SNP has the ability to affect gene function by regulating mRNA splicing [33], stability [34] and protein translation [35]. However, synonymous SNP is not a preferred polymorphism to be further validated. Therefore, the utilization of non-synonymous SNPs is important to prioritize their involvement in the biosynthetic pathway and to determine the effect of non-synonymous SNPs to phenotypic expression.

SNPs in the intron region are able to shift the gene splicing or regulate the transcript level by changing the binding sites of miRNA [36]. Lower number of SNPs in untranslated region (UTR) has demonstrated that mutation in this region can change local mRNA structure and will affect the translation process [36]. SNPs in the UTR region are conserved and consist of binding sites for proteins or antisense RNAs that able to modulate transport, RNA stability, cellular localization, expression level and translation [37].

Functional effect of SNPs in the flavonoid related genes can affect the expression and regulation of structural genes and transcription factor during the flavonoid biosynthesis process. Previous study has demonstrated that genetic variation caused by SNP can influenced the variation and accumulation of metabolites in the biosynthetic pathway [19, 20]. By selecting SNPs with specific functional effect such as non-synonymous and deleterious, it can be effectively used as a molecular marker in the development of new and improved rice variety with agronomic traits of interest.

However, it is a challenge to depend only on SNPs to reveal the interaction and relationship that occur in the multiple levels of biological mechanisms [38], especially in the qualitative and complex traits (i.e. abiotic stress, quality, yield).

Therefore, numerous studies have been performed to identify and evaluate the functional effects of SNPs through multi-omics data integration. For instance, the integration of whole genome and transcriptome data offers opportunities to highlight the expressed SNPs [39] and to better understand the biological processes and mechanisms underlying the pigmented rice varieties. Previous study has performed integration of SNPs with gene co-expression network to identify the causal genes and causal SNPs that might be responsible in the mechanism of blast disease resistance [40], salt tolerance [41] and amylose content [42]. Throughout these integration processes, several bioinformatics tools have been applied to identify the putative SNPs and to annotate the SNPs into their functional effects.

The identification of SNPs in the flavonoid related genes started with mining of SNPs from the genome and transcriptome datasets of *O. sativa indica* cv. Bali, PH9, MRM16, MRQ100, MR297 and MRQ76 [43, 44]. Bali, PH9, MRM16 and MRQ100 are pigmented rice varieties that contain high antioxidant contents. MR297 and MRQ76 are white rice varieties with medium tolerance to diseases. **Figure 1** shows the bioinformatic workflow to identify SNPs from the genome and transcriptome datasets. The first step in SNPs identification is to map the sequence reads onto selected reference genome sequences; in this case it was Nipponbare. It was chosen because of it was well-assembled and annotated. The genome and transcriptome of reads mapping were individually aligned using different mapping tools such as BWA [45] for genome reads mapping and TopHat2 [46] for transcriptome reads mapping.

Then PICARD version 0.7.12 [47] was used to add and replace the read groups, marking the duplicate reads and fixing mate information on the mapped reads in order to obtain the high-quality SNPs during SNPs calling. This process is known as post-processing and it is a standard step used to identify potential SNPs from the genome and transcriptome datasets. Once this process is completed, GATK version 3.6 [47] was used in the SNPs calling process and this is a crucial step in SNPs discovery as it helps in obtaining high-quality SNPs as well as reducing false-positive SNPs. In our case, the parameters used are as follow:

i. mapping quality (MQ) more than and equal to 30 (MQ > = 30);

ii. variant quality more than and equal to 50 (VQ > =50);

iii. number of supporting reads for every base (DP) more than and equal to 10 (DP > = 10)

All SNPs obtained from the filtering process were annotated using several annotation tools such as SnpEff version 4.1 [48], Variant Effect Predictor (VEP) [49], CooVar [50] and Annovar [51] for their intergenic, genic, coding region, UTR, intron, synonymous and non-synonymous. Most of these tools can be locally installed to ease the users who perform large-number of SNPs annotation. Those SNPs that were annotated as genic were then filtered using R packages (i.e. dplyr, tidyr, sqldf) for the identification of SNP position, chromosome, allele, gene identifier and SNPs effect. This information was then used in matching the SNPs in genome and transcriptome datasets as well as screening all SNPs that reside in the flavonoid related genes.

Comparative SNPs analysis of six rice varieties were performed to investigate the uniqueness, differences and similarities in the potential SNPs. Currently several rice SNP databases are available to providing the information on SNPs that were mined from various rice varieties and sub-specie such as Rice SNP-Seek [52],

Rice VarMap [53], IC4R [54], and Ensembl Plants Variation database [55]. Comparative SNPs analysis on SNP databases can provide the SNPs information and usability in various rice varieties and sub-species.

SNP occurs in transcription factor often lead to altering or loss of function of key pathway enzymes that are required to regulating the production of anthocyanin [56]. Hence, it could affect the expression levels of flavonoid. Previous research has investigated that mutations were accumulated during the domestication process, which suggests the presence of agronomically valuable genes in landraces as well as in wild relative [15]. Hence, SNPs in transcription factor, such as *Kala*4 and *Rc,* can be prioritized for further integration with genes co-expression network.

3.3 Gene co-expression network analysis from transcriptome data

A gene co-expression network analysis was performed to correlate gene and phenotypic expressions, to infer the function of unknown genes and to identify the key regulatory networks in biosynthetic processes [57, 58]. The principle used in the gene co-expression network shows that genes that cluster in the same network represent similar biological process [58]. To date, the increasing number of genes co-expression network databases such as ATTED, STRING, RED facilitate the exploitation of co-expressed genes from different conditions in various crops [59]. In the co-expressed gene databases, the gene identifier or gene name can be used as a query to search for the co-expressed genes of interest. Parameter, such as maximum number of interactors = 0.1 and confidence score cut-off = 0.40 are used to select the significant co-expressed genes. Most of the co-expressed databases obtained the information from microarray gene expression and transcriptome datasets from public databases such as NCBI Gene Expression Omnibus (NCBI GEO) [60] and Sequence Read Archive (SRA) [61].

Gene co-expression network analysis was performed using the expression values of genes in fragments per kilobase of transcript per million mapped reads (FPKM) as an input data. Pearson correlation coefficient (PCC) value was used to measure the co-expression correlation between paired genes in the network. The PCC score represents the confidence level describing the association of the two genes whether they are functionally associated. In this case, the FPKM value with more than and equal to 0.1 (FPKM > = 0.1) was used to perform a gene co-expression network analysis. Pearson Correlation Coefficient (PCC) in corr R package [62] was used to measure the correlation between paired genes in the network. The cut-off value of PCC more than and equal to 0.7 (PCC > = 0.7) was used for selection of co-expressed genes. The flavonoid related genes were selected as guided-genes to identify their interactors. **Figure 1** shows the analysis workflow to perform a gene co-expression network analysis.

Cytoscape v3.7 was used to construct and visualize the gene co-expression network. The Network Analyzer plugin was used to calculate the degree connectivity of each nodes and edges. ShinyGO version 0.6.1 [63] was used in gene ontology and pathway enrichment analysis to elucidate the biological processes and molecular function in clusters of network.

4. Data integration

4.1 Integration of SNPs with co-expressed genes

There are several publications on the bioinformatics approaches on the functional effect of SNPs where by assessing their effect on the functional site of protein

structure [64, 65] and integrate with biological pathway [66]. In this study, SNP, especially with non-synonymous effect and deleterious impact is better appreciated to study its impact through the pathway-network analysis coupled with a different omics data type.

SNPs can be integrated with omics data type to rank the high-potential SNPs and to highlight the causal genes for development of genetic markers and functional genomics studies. The integration of SNPs and co-expressed genes through construction of biological network offers a comprehensive interpretation of genetic variation at the biological system level. Integrative approach links biologically meaningful sets of genes to reveal the molecular basis of trait variation.

In this study, integration of SNPs and co-expressed genes was performed to prioritize the functional SNPs that can be suggested as a candidate for the development of molecular markers and to highlight the causal genes that might contribute in the flavonoid biosynthesis process. The co-expressed genes can be suggested as a functional target for functional validation in the future study to unravel their potential in rice breeding improvement.

Integration of SNPs and co-expressed genes was performed using Cytoscape v3.7 [67]. Flavonoid biosynthetic pathway templates in BioPAX, KGXML and SIF formats was retrieved from KEGG, Plant Reactome and RiceCyc databases. To merge the SNP and co-expressed genes, gene identifier (ID) was used as a matching identity. Every gene in the network consists of gene ID, chromosome, start and stop position based on physical genome coordinates (bp). Similarly, SNPs consists of the position in genome locations (bp) and gene ID.

Analysis on network integration between SNPs and co-expressed genes highlight that co-expressed genes can be integrated by multiple numbers of SNPs and will reveal, which SNPs appear to play an essential role in the flavonoid biosynthesis process. The co-expressed genes connected to SNPs can be prioritized as candidate genes. The high false-positive rate in SNPs also can be reduced by incorporating putative functional co-expressed genes information [68]. Several biological questions can be asked from the integration of SNPs and co-expressed genes into pathway-network, such as i) how many functional SNPs and co-expressed genes important to the expression of black and red pigmentation; ii) any regulatory genes regulate the biosynthetic pathway?; iii) and which biological process are underlying this trait.

4.2 Pathway-network analysis

There are different ways in bioinformatics approaches that could be applied to the pathway-network analysis. In pathway-network analysis, the description of connected genes can be interpreted into biologically meaningful information and provide insights into biological processes, molecular function and cellular components. This analysis is known as gene ontology (GO) enrichment and pathway enrichment analysis.

To date, several bioinformatic tools are available to perform GO and pathway enrichment analysis in the network. For example, ClueGO [69] and BiNGO [70]. Both plugins are available in the Cytoscape and are user-friendly. Statistical values such as Hypergeometric testing and Bonferroni method is used to calculate the p-value. Parameter such as p-value less and equal than 0.05 (p-value <= 005) and a minimum number of mapping entries > = 2 can be used to select the significant or enrich genes in the pathway-network.

Tables 1 and **2** provide list of bioinformatics tools and databases that are used for data mining and data integration in search for potential SNPs involved in the flavonoid biosynthesis.

Bioinformatic tools	Description	References
BLASTX	Sequence similarity search analysis	[22]
eggNog	Annotation of orthologous and paralogous sequences	[23]
BWA	Genome mapping	[45]
TopHat2 v2.3	Transcriptome mapping	[46]
Picard v0.7.12	Post mapping processes	[47]
GATK v3.6 (HaplotypeCaller)	SNPs discovery	[47]
SnpEff v4.1	SNPs annotation	[48]
R package (corrr)	Gene co-expression analysis	[62]
ShinyGO	Gene ontology enrichment analysis	[63]
Cytoscape v3.7	Data integration	[66]
Cytoscape v3.7 plugin ClueGO	Gene ontology enrichment analysis	[68]
R packages (dplyr, tidyr, sqldf)	Data cleaning and filtering	[71] [72] [73]

Table 1.
Summary of the bioinformatic tools used in search for potential SNPs involved in the flavonoid biosynthetic pathways.

Bioinformatic databases	Descriptions	URL	References
Rice Annotation Project Database (RAP-DB)	Is a rice genome database that has been developed by the International Rice Genome Sequencing. Information provided are genome sequences, chromosom, gene annotation dan description.	https://rapdb.dna.affrc.go.jp/	[74]
Kyoto Encyclopedia of Genes and Genomes (KEGG)	A pathway database that provides biological information related to genes, proteins, enzymes and pathways involve in biological systems.	https://www.kegg.jp/	[75]
PlantReactome	A plant pathway database that provides user genes, proteins, enzymes and reactions involve in the specific biological systems.	https://plantreactome.gramene.org/	[76]
RiceCyc	A rice metabolic pathway database that has been developed to provide predicted biochemical pathways in rice. Several biological information, such as genes, proteins, enzymes and reactions have been displayed in diagram and all the data can be downloaded by user.	http://pathway.gramene.org/gramene/ricecyc.shtml	[77]

Table 2.
Summary of the biological databases used in search for potential SNPs involved in the flavonoid biosynthetic pathways.

5. Repository of omics data

Continuously increased number of biological data is in need for a database to systematically store, organize and manage them. To date, several rice databases are available for the application in rice breeding programme, such as rice genome

Figure 2.
The homepage of MyNutRice*Base (http://www.mynutricebase.org).*

database [74], SNP databases [53, 54, 78] and pathway databases [75, 76]. Each of these databases have their uniqueness and specific target users.

In this study, the flavonoid related gene, co-expressed gene and SNP are stored in a one-stop database that is specifically developed as a genetics and genomics repository to keep all information related to nutritional traits in rice known as *MyNutRice*Base (http://www.mynutricebase.org) (**Figure 2**). It provides a platform for data mining of SNPs and genes, data visualization and sequence similarity search analysis. *MyNutRice*Base aims to accelerate the genomics and genetic analysis by enabling the rice geneticist and breeders to mine and export the biological information for the application in rice breeding improvement.

6. Conclusions

To date, several bioinformatics tools are available for the researcher to use and connect in an analysis pipeline. Criteria and parameter are the essential part that always carefully revised and look into while performing the bioinformatics analysis. This review highlights bioinformatics workflow used in the identification of SNPs in genomic and transcriptomic data, gene co-expression network analysis, omics data integration. This result facilitates the interpretation of SNPs

into comprehensive biological information with the identification of potential. By integrating these two types of data, it may shed some light on the roles of various genes and SNPs that may play an essential role in the accumulation of flavonoid content in rice.

Acknowledgements

This work was carried out at the Centre for Bioinformatics Research, Institute of Systems Biology (INBIOSIS), Universiti Kebangsaan Malaysia and Malaysian Agricultural Research & Development Institute (MARDI). The open access publishing fees are funded by GP-2020-K007217 and GP-2019-K021204 grants.

Conflict of interest

The authors declare no conflict of interest.

Author details

Rabiatul-Adawiah Zainal-Abidin[1,2] and Zeti-Azura Mohamed-Hussein[1,3*]

1 Institute of Systems Biology (INBIOSIS), Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

2 Malaysian Agricultural Research and Development Institute (MARDI), Serdang, Malaysia

3 Department of Applied Physics, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

*Address all correspondence to: zeti.hussein@ukm.edu.my

References

[1] Wang L, Zhong M, Li X, Yuan D, Xu Y, Liu H, et al. The QTL controlling amino acid content in grains of rice (Oryza sativa) are co-localized with the regions involved in the amino acid metabolism pathway. Mol Breed. 2008;21(1):127-37. DOI: 10.1007/s11032-007-9141-7.

[2] Ye X, Salim A-B, Kloti A, Zhang J, Lucca P, Beyer P, et al. Engineering the Provitamin A (Beta-Carotene) Biosynthetic Pathway into (Carotenoid-Free) Rice Endosperm. Science (80). 2000;287:303-5. DOI: 10.1126/science.287.5451.303.

[3] Stangoulis JCR, Huynh BL, Welch RM, Choi EY, Graham RD. Quantitative trait loci for phytate in rice grain and their relationship with grain micronutrient content. Euphytica. 2007;154(3):289-94. DOI: 10.1007/s10681-006-9211-7.

[4] Maeda H, Yamaguchi T, Omoteno M, Takarada T, Fujita K, Murata K, et al. Genetic dissection of black grain rice by the development of a near isogenic line. Breed Sci. 2014;64(2):134-41. DOI: 10.1270/jsbbs.64.134.

[5] Aung MS, Masuda H, Kobayashi T, Nakanishi H, Yamakawa T, Nishizawa NK. Iron Biofortification of Myanmar Rice. Front Plant Sci. 2013;4:1-14. DOI: 10.3389/fpls.2013.00158.

[6] Nachimuthu VV, Robin S, Sudhakar D, Rajeswari S, Raveendran M, Subramanian KS, et al. Genotypic variation for micronutrient content in traditional and improved rice lines and its role in biofortification programme. Indian J Sci Technol. 2014;7(9):1414-25. DOI: 10.1.1.1028.9732.

[7] Anukul N, Ramos RA, Mehrshahi P, Castelazo AS, Parker H, Diévart A, et al. Folate polyglutamylation is required for rice seed development. Rice. 2010;3(2-3):181-93. DOI: 10.1007/s12284-010-9040-0.

[8] Dipti SS, Bergman C, Indrasari SD, Herath T, Hall R. The potential of rice to offer solutions for malnutrition and chronic diseases. Rice. 2012;5:1-18. DOI: 10.1186/1939-8433-5-16.

[9] Petrussa E, Braidot E, Zancani M, Peresson C, Bertolini A, Patui S, et al. Plant flavonoids-biosynthesis, transport and involvement in stress responses. Int J Mol Sci. 2013;14(7):14950-73. DOI: 10.3390/ijms140714950.

[10] Tohge T, De Souza LP, Fernie AR. Current understanding of the pathways of flavonoid biosynthesis in model and crop plants. J Exp Bot. 2017;68(15):4013-28. DOI: 10.1093/jxb/erx177.

[11] Lepiniec L, Debeaujon I, Routaboul J-M, Baudry A, Pourcel L, Nesi N, et al. Genetics and Biochemistry of Seed Flavonoids. Annu Rev Plant Biol. 2006;57(1):405-30. DOI: 10.1146/annurev.arplant.57.032905.105252.

[12] Oikawa T, Maeda H, Oguchi T, Yamaguchi T, Tanabe N, Ebana K, et al. The birth of a black rice gene and its local spread by introgression. Plant Cell. 2015; 27(9):2401-2414. DOI: 10.1105/tpc.15.00310.

[13] Sweeney MT, Thomson MJ, Pfeil BE, McCouch S. Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. Plant Cell. 2006;18(2):283-94. DOI: 10.1105/tpc.105.038430.

[14] Sakamoto W, Ohmori T, Kageyama K, Miyazaki C, Saito A, Murata M, et al. The Purple leaf (Pl) locus of rice: the Pl(w) allele has a complex organization and includes

two genes encoding basic helix-loop-helix proteins involved in anthocyanin biosynthesis. Plant Cell Physiol. 2001;42(9):982-91. DOI: 10.1093/pcp/pce128.

[15] Saitoh K, Onishi K, Mikami I, Thidar K, Sano Y. Allelic diversification at the C (*OsC1*) locus of wild and cultivated rice: Nucleotide changes associated with phenotypes. Genetics. 2004;168(2):997-1007. DOI: 10.1534/genetics.103.018390.

[16] Goufo P, Trindade H. Rice antioxidants: phenolic acids, flavonoids, anthocyanins, proanthocyanidins, tocopherols, tocotrienols, γ -oryzanol, and phytic acid. Food Sci Nutr. 2014;2(2):75-104. DOI: 10.1002/fsn3.86.

[17] Shih CH, Chu H, Tang LK, Sakamoto W, Maekawa M, Chu IK, et al. Functional characterization of key structural genes in rice flavonoid biosynthesis. Planta. 2008;228(6):1043-54. DOI: 10.1007/s00425-008-0806-1.

[18] Dong X, Chen W, Wang W, Zhang H, Liu X, Luo J. Comprehensive profiling and natural variation of flavonoids in rice. J Integr Plant Biol. 2014;56(9):876-86. DOI: 10.1111/jipb.12204.

[19] Clotault J, Peltier D, Soufflet-Freslon V, Briad M, Geoffriau E. Differential selection on carotenoid biosynthesis genes as a function of gene position in the metabolic pathway: A study on the carrot and dicots. PLoS One. 2012;7(6):1-13.

[20] Manrique-Carpintero NC, Tokuhisa JG, Ginzberg I, Holliday J a, Veilleux RE. Sequence diversity in coding regions of candidate genes in the glycoalkaloid biosynthetic pathway of wild potato species. Genes|Genomes|Genetics. 2013;3(9):1467-79. 10.1371/journal.pone.0038724.

[21] Jin L, Xiao P, Lu Y, Shao Y, Shen Y, Bao J. Quantitative Trait Loci for Brown Rice Color , Phenolics , Flavonoid Contents , and Antioxidant Capacity in Rice Grain. Cereal Chem. 2009;86(6):609-15. DOI: 10.1007/s00122-010-1505-4.

[22] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):3-10.DOI:10.1016/S0022-2836(05)80360-2.

[23] Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2018;47(1): 9-14. DOI: 10.1093/nar/gky1085.

[24] Qu C, Zhao H, Fu F, Wang Z, Zhang K, Zhou Y, et al. Genome-Wide Survey of Flavonoid Biosynthesis Genes and Gene Expression Analysis between Black- and Yellow-Seeded *Brassica napus*. Front Plant Sci. 2016;7(1). DOI:10.3389/fpls.2016.01755.

[25] Dequigiovanni G, Ritschel PS, Maia JDG, Quecini V. In silico SNP detection for anthocyanin metabolism genes in Vitis. Acta Hortic. 2014;1046:341-8. DOI: 10.17660/ActaHortic.2014.1046.46.

[26] Singh N, Singh B, Rai V, Sidhu S, Singh AK, Singh NK. Evolutionary Insights Based on SNP Haplotypes of Red Pericarp, Grain Size and Starch Synthase Genes in Wild and Cultivated Rice. Front Plant Sci. 2017;8(1):1-9. DOI:10.3389/fpls.2017.00972.

[27] McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, et al. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc Natl Acad Sci U S A. 2009;106(30):12273-8. DOI: 10.1073/pnas.0900992106.

[28] Zhao K, Tung C, Eizenga GC, Wright MH, Ali ML, Price AH, et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat Commun. 2011;2(467):1-10. DOI: 10.1038/ncomms1467.

[29] Edwards D, Forster JW, Chagné D, Batley J. What are SNPs? In: Association mapping in plants. New York: Springer; 2005. p. 41-52. DOI: 10.1007/978-0-387-36011-9_3.

[30] Kono TJY, Fu F, Mohammadi M, Hoffman PJ, Liu C, Stupar RM, et al. The Role of Deleterious Substitutions in Crop Genomes. Mol Biol Evol. 2016;33(9):2307-17. DOI: 10.1093/molbev/msw102.

[31] Huq A, Akter S, Nou S, Kim HT, Jung YJ, Kang KK. Identification of functional SNPs in genes and their effects on plant phenotypes. J Plant Biotechnol. 2016;43(1):1-11. DOI: 10.5010/JPB.2016.43.1.1.

[32] Günther T, Schmid KJ. Deleterious amino acid polymorphisms in *Arabidopsis thaliana* and rice. Theor Appl Genet. 2010;121:157-68. DOI: 10.1007/s00122-010-1299-4.

[33] Pagani F, Raponi M, Baralle FE. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci. 2005;102(18):6368-72. DOI: 10.1073/pnas.0502288102.

[34] Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar S V., et al. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science (80). 2007;315(5811):525-8. DOI: 10.1126/science.1135308.

[35] Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet. 2011;12(10):683-91. DOI: 10.1038/nrg3051.

[36] Tatarinova T V, Chekalin E, Nikolsky Y, Bruskin S, Chebotarov D, McNally KL, et al. Nucleotide diversity analysis highlights functionally important genomic regions. Sci Rep. 2016;6(1):1-12. DOI: 10.1038/srep35730.

[37] Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: Synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res. 2013;41(4):2073-94. DOI: 10.1093/nar/gks1205.

[38] Yuan JS, Galbraith DW, Dai SY, Griffin P, Stewart CN. Plant systems biology comes of age. Trends Plant Sci. 2008;13(4):165-71. DOI: 10.1016/j.tplants.2008.02.003.

[39] Seol Y, Won SY, Shin Y, Lee J, Chun J, Kim Y, et al. A Multilayered Screening Method for the Identification of Regulatory Genes in Rice by Agronomic Traits. Evol Bioinforma. 2016;12:253-62. DOI: 10.4137/EBO.S40622.

[40] Ficklin SP, Feltus FA. A Systems-Genetics Approach and Data Mining Tool to Assist in the Discovery of Genes Underlying Complex Traits in *Oryza sativa*. PLoS One. 2013;8(7):1-7. DOI: 10.1371/journal.pone.0068551.

[41] Wang J, Chen L, Wang Y, Zhang J, Liang Y, Xu D. A Computational Systems Biology Study for Understanding Salt Tolerance Mechanism in Rice. PLoS One. 2013;8(6):1-12. DOI: 10.1371/journal.pone.0064929.

[42] Butardo VM, Anacleto R, Parween S, Samson I, de Guzman K, Alhambra CM, et al. Systems Genetics Identifies a Novel Regulatory Domain of Amylose Synthesis. Plant Physiol. 2017;173(1):887-906. DOI:10.1104/pp.16.01248.

[43] Zainal-Abidin R-A, Zainal Z, Mohamed-Hussein Z-A, Abu-Bakar N, Ab Razak S, Simoh S, et al. RNA-seq data from whole rice grains of pigmented and non-pigmented Malaysian rice varieties. Data Br. 2020;30:105432. DOI: 10.1016/j.dib.2020.105432.

[44] Zainal-Abidin R-A, Zainal Z, Mohamed-Hussein Z-A, Sew YS, Simoh S, Ab Razak S, et al. Data on genome resequencing of pigmented and non-pigmented Malaysian rice varieties. Data Br. 2020;31:105806. DOI: 10.1016/j.dib.2020.105806.

[45] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. DOI: 10.1093/bioinformatics/btp324.

[46] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2 : accurate alignment of transcriptomes in the presence of insertions , deletions and gene fusions. Genome Biol. 2013;14(4):1-13. DOI: 10.1186/gb-2013-14-4-r36.

[47] Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;1-33. DOI: 10.1002/0471250953.bi1110s43.

[48] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80-92. DOI: 10.4161/fly.19695.

[49] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):1-14. DOI: 10.1186/s13059-016-0974-4.

[50] Vergara IA, Frech C, Chen N. CooVar: Co-occurring variant analyzer. BMC Res Notes. 2012;5(615):1-7. DOI: 10.1186/1756-0500-5-615.

[51] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):1-7. DOI: 10.1093/nar/gkq603.

[52] Alexandrov N, Tai S, Wang W, Mansueto L, Palis K, Fuentes RR, et al. SNP-Seek database of SNPs derived from 3000 rice genomes. Nucleic Acids Res. 2015;43:1023-7. DOI: 10.1093/nar/gku1039.

[53] Zhao H, Yao W, Ouyang Y, Yang W, Wang G, Lian X, et al. RiceVarMap: A comprehensive database of rice genomic variations. Nucleic Acids Res. 2015;43(1):18-22. DOI: 10.1093/nar/gku894.

[54] Zhang Z, Hu S, He H, Zhang H, Chen F, Zhao W, et al. Information Commons for Rice (IC4R). Nucleic Acids Res. 2016;44:1172-80. DOI: 10.1093/nar/gkv1141.

[55] Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, et al. Ensembl variation resources. Database. 2018;2018:1-12. DOI: 10.1093/database/bay119.

[56] Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, et al. Co-expression tools for plant biology: Opportunities for hypothesis generation and caveats. Plant, Cell Environ. 2009;32(12):1633-51. DOI: 10.1111/j.1365-3040.2009.02040.x.

[57] Li Y, Pearl SA, Jackson SA. Gene Networks in Plant Biology: Approaches in Reconstruction and Analysis. Trends Plant Sci. 2015;20(10):664-75. DOI: 10.1016/j.tplants.2015.06.013.

[58] van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene

co-expression analysis for functional classification and gene–disease predictions. Briefing in Bioinformatics. 2017;1-18. DOI: 10.1093/bib/bbw139.

[59] Obayashi T, Aoki Y, Tadaka S, Kagaya Y, Kinoshita K. ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index. Plant Cell Physiol. 2017;59(1):1-7. DOI: 10.1093/pcp/pcx191/4690683.

[60] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: Archive for functional genomics data sets - Update. Nucleic Acids Res. 2013;41(1):91-95. DOI: 0.1093/nar/gks1193.

[61] Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011;39(1):19-21. DOI: 10.1093/nar/gkq1019.

[62] Kuhn M, Jackson S and Cimentada J. corrr: Correlations in R. R package version 0.4.2. (2020). https://CRAN.R-project.org/package=corrr.

[63] Ge SX, Jung D. ShinyGO: a graphical enrichment tool for animals and plants. bioRxiv. 2018;(315150):2. DOI: 10.1101/315150v1.

[64] Bhardwaj A, Dhar YV, Asif MH, Bag SK. In Silico identification of SNP diversity in cultivated and wild tomato species: Insight from molecular simulations. Sci Rep. 2016;6(1):1-13. DOI: doi.org/10.1038/srep38715.

[65] Sholikhah A, Khasna EN, Dahlia, Listyorini D. Polymorphism of gene encoding granule bound starch synthase i (GBSSI) involved in starch biosynthesis in local rice from Banyuwangi. AIP Conf Proc. 2018;2002. DOI: 10.1063/1.5050100.

[66] Cirillo E, Kutmon M, Hernandez MG, Hooimeijer T, Adriaens ME, Eijssen LMT, et al. From SNPs to pathways: Biological interpretation of type 2 diabetes (T2DM) genome wide association study (GWAS) results. PLoS One. 2018;13(4):1-19. DOI: 10.1371/journal.pone.0193515.

[67] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research. 2003 Nov;13(11):2498-504. doi: 10.1101/gr.1239303.

[68] Schaefer RJ, Michno J-M, Jeffers J, Hoekenga O, Dilkes B, Baxter I, et al. Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. Plant Cell. 2018;30(12):2922-42. DOI: 10.1105/tpc.18.00299.

[69] Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25(8):1091-3. DOI: 0.1093/bioinformatics/btp101.

[70] Maere S, Heymans K, Kuiper M. BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics. 2005;21(16):3448-9. DOI: 10.1093/bioinformatics/bti551.

[71] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller. dplyr: A Grammar of Data Manipulation. R package version 0.7.6. (2018). https://CRAN.R-project.org/package=dplyr.

[72] Hadley Wickham and Lionel Henry. tidyr: Tidy Messy Data. R package version 1.0.2. (2020). https://CRAN.R-project.org/package=tidyr

[73] G. Grothendieck. sqldf: Manipulate R Data Frames Using SQL. R package version 0.4-11. (2017). https://CRAN.R-project.org/package=sqldf.

[74] Matsumoto T, Wu J, Itoh T, Numa H, Antonio B, Sasaki T. The Nipponbare genome and the next-generation of rice genomics research in Japan. Rice. 2016;9(33). DOI: 10.1186/s12284-016-0107-4.

[75] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:457-62. DOI: 10.1093/nar/gkv1070.

[76] Naithani S, Preece J, D'Eustachio P, Gupta P, Amarasinghe V, Dharmawardhana PD, et al. Plant Reactome: A resource for plant pathways and comparative analysis. Nucleic Acids Res. 2017;45(1):1029-1039. DOI: 10.1093/nar/gkw932.

[77] Dharmawardhana P, Ren L, Amarasinghe V, Monaco M, Thomason J, Ravenscroft D, McCouch S, Ware D, Jaiswal P. A genome scale metabolic network for rice and accompanying analysis of tryptophan, auxin and serotonin biosynthesis regulation under biotic stress. Rice. 2013;6(1):1-15. DOI: 10.1186/1939-8433-6-15.

[78] Hawkins C, Caruana J, Schiksnis E, Liu Z. Genome-scale DNA variant analysis and functional validation of a SNP underlying yellow fruit color in wild strawberry. Sci Rep. 2016;6(July):1-11. DOI: 10.1038/srep29017.