Construction of transcriptional regulatory network proposes bZIP transcription factor controlling Rubisco genes in cassava

Bandit Khampoosa[1], Somkid Bumee[2], Treenut Saithong[1,2,3], Malinee Suksangpanomrung[4] and Saowalak Kalapanulak[1,2,3,*]

[1]*Bioinformatics and Systems Biology Program, School of Bioresources and Technology and School of Information Technology, King Mongkut's University of Technology Thonburi, Bang Khun Thian, Bangkok, Thailand, 10150*
[2]*Systems Biology and Bioinformatics Research Laboratory, Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi, Bangkhuntien, Bangkok, Thailand, 10150*
[3]*School of Bioresources and Technology, King Mongkut's University of Technology Thonburi ,Bang Khun Thian, Bangkok, Thailand, 10150*
[4]*National Center for Genetic Engineering and Biotechnology (BIOTEC), Pathumthani, Thailand 12120,*
**e-mail saowalak.kal@kmutt.ac.th*

Abstract

Cassava (*Manihot esculenta* Crantz) is one of the most important carbohydrate resources for human beings around the world, especially in Africa and Asia. Its roots contain higher starch content, up to 90% dry weight, than other starchy crops. Because of its advantages, the starch biosynthesis pathway and its transcriptional regulation are of great interest to investigate to gain more understanding about the plant. In this work, the bioinformatics approach called template-based method has been applied to construct the transcriptional regulatory network (TRN) of starch metabolism in cassava. TRN of *Arabidopsis thaliana (A. thaliana)*, composed of 11,354 interactions from 67 transcription factors (TFs) and its targets, was used as a template for inferring the TRN of cassava. Transcription factor binding sites of predicted TF on the promoter region of their target genes have been verified through PlantPAN database. The results show that two cassava genes (cassava4.1_017170m.g and cassava4.1_017243m.g), which encode for ribulose-bisphosphate carboxylase (EC4.1.1.39), could be controlled by transcription factor LONG HYPOCOTYL5 (HY5) (cassava4.1_017720m.g) classified in the basic-leucine zipper (bZIP) transcription factor family. Moreover, the gene expression profiles of TF and target genes under cold stress in cassava apical shoot tissue were then shown the correlation between them via Pearson's correlation coefficient. A high correlation of one gene pair's expression profiles was implied the regulation of HY5 to Rubisco gene (cassava4.1_017243m.g). It may contribute to gain a new cassava cultivar with high starch yield via increasing photosynthesis efficiency.

Keywords: bZIP, cassava, Rubisco, transcription factor, transcriptional regulatory network

Introduction

Cassava is a starchy root crop in the *Euphorbiaceae* family of *Dicotyledonae* class. It is an important source of food and calories in the world's tropics because its roots contain high starch content and proper starch properties such as odorless, paste clarity, and stickiness. Cassava is not only a source for human food, but it is also a resource for diverse industries, including animal feed and biofuel production. Thailand is the largest exporter of raw cassava to the world market. To address the growing demand for starch products in food and other industrial applications, the understanding of starch metabolism and its transcriptional regulation are crucial to gain the ability to improve starch production and modification in cassava plant.

Although cassava is one of the important starchy root crops, current understanding of starch biosynthesis pathway relies on biological investigation in the model plant, *A. thaliana.* Since cassava genome was released in 2011 (Goodstein et al., 2011), a high-quality pathway of cassava starch metabolism was reconstructed by using the genomics information from plant templates based on comparative genomic protocol (Saithong et al., 2013). The pathways cover the main processes including the Calvin cycle, sucrose synthesis, and storage starch biosynthesis. Next, transcriptome data was also integrated on the starch metabolism to gain more understanding about gene function in the particular conditions or plant tissues (Wanatsanan et al., 2012). However, the regulation of gene in the transcription level and phenotypic responses from various perturbations in cassava are still less.

With regard to the gene regulation, transcription factors (TFs) are the regulatory proteins interacting with the specific sites on the promoters of target genes to activate or repress the expression of target gene. Several computational approaches were then applied to identify transcriptional regulatory network (TRN) in both single- and multiple-cell organisms (Babu et al., 2008). Template-based method was simple method applied to infer the orthologous TRN of an interested organism by using known TRN of another organism. Liu and co-workers (Liu et al., 2008) reconstructed TRN of *Shewanella oneidensis* by using known TRN of *Esherichia coli* as the template. Moreover, reverse engineering approaches were successfully used to infer co-regulated genes from gene expression patterns in microarray data. Co-expression networks of plant model and its functional modules of starch metabolism were reconstructed by using 1,094 microarrays based on the Pearson correlation (Mao et al., 2009). The TRNs of starch metabolism in *A. thaliana* leaves under diurnal cycle (Ingkasuwan et al., 2012) and constant light conditions (Bumee et al., 2013) were discovered based on graphical Gaussian models (GGM). However, the performance of the inferred TRN based on co-expression networks depends on number of microarray datasets (Huang et al., 2007) and size of data points in each microarray dataset (Vijender et al., 2012). Predicting *cis*-regulatory elements was another method for inferring regulatory network (Janky et al., 2009). This makes use of known characterized transcription factor binding sites (TFBSs) to predict regulatory interactions between TFs and their targets in other organisms. Unfortunately in case of cassava, a few short time series microarray datasets are available (Li et al., 2010; Yang et al., 2011). Therefore, inferring TRN of cassava by using co-expression network is limited.

This work aims to reconstruct TRN of starch metabolism in cassava by taking the advantages of the template-based method and *cis*-regulatory elements analysis. The well-known TRN of *A. thaliana* (Davuluri et al., 2003) was used as a template for inferred TRN in cassava. Seventy-nine of 11,354 known interactions between 52 genes and 11 transcription factors were identified as TRN of starch metabolism in *A. thaliana*. Three orthologous genes in cassava were then identified based on TRN of starch metabolism in *A. thaliana*. Two cassava genes (cassava4.1_017170m.g and cassava4.1_017243m.g) encoding for ribulose-bisphosphate carboxylase (Rubisco) were controlled by a putative transcription factor LONG HYPOCOTYL5 (HY5) classified in the basic-leucine zipper (bZIP) transcription factor family. The binding possibility of TFs on the upstream region of these two target genes was verified through *cis*-regulatory element analysis based on PlantPAN database. Moreover, the regulation function of TF to targets was then confirmed via the co-expression behavior in cassava apical shoots under cold stress (An et al., 2012). A correlation of each gene pair's expression profiles was implied the regulation of HY5 to Rubisco genes. The results provide the high confidence for experimentalist to confirm this regulation in cassava plant. It may contribute to obtain a new cultivar with high starch yield via the improvement of

photosynthesis efficiency in cassava.

Methodology

This work is divided into three main parts: 1) collecting regulatory interactions between TFs and target genes functioning in starch metabolism in *A. thaliana* 2) inferring TRN of cassava by using the template-based method, and 3) verifying the binding possibility between predicted TFs and target genes via *cis*-regulatory element analysis (Figure 1).

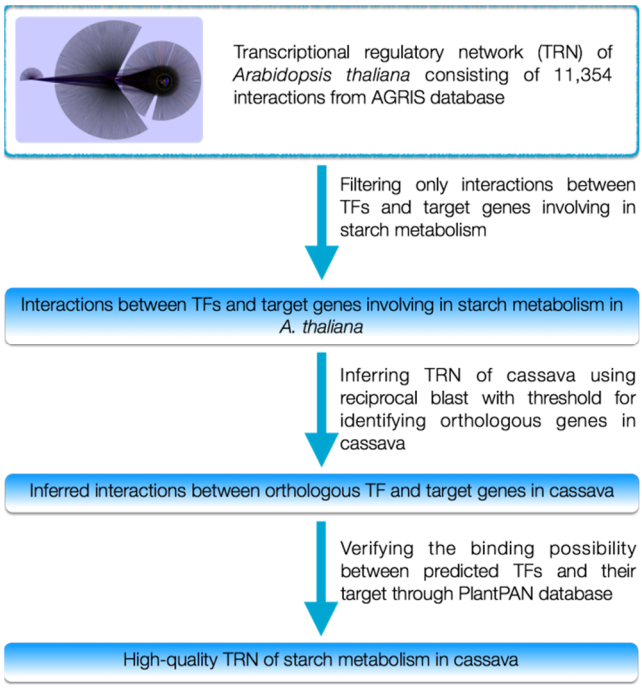

Figure 1: The overall methodology

Collecting interactions between TFs and targets functioning in starch metabolism in *A. thaliana*.

TRN of *A. thaliana* used as a template for reconstructing TRN of starch metabolism in cassava was downloaded from The Arabidopsis Gene Regulatory Information Server (AGRIS) database (Davuluri et al., 2003). The enzymatic genes functioning in starch metabolism including Calvin cycle, sucrose synthesis, and storage starch biosynthesis were identified based on comparative approach in the previous work (Saithong et al., 2013).

Inferring TRN of cassava by using template-based method

The enzymatic genes that function in starch metabolism were selected from TRN of *A. thaliana* resulted in TF-target genes pairs. TFs and enzymatic genes in *A. thaliana* were used to identify orthologous genes in cassava by comparing all protein sequences of these genes with all proteins in the cassava genome gathered from Phytozome database (Goodstein et al., 2011). Othologous enzymatic genes in cassava was identified by reciprocal BLASTp protocol (Saithong et al., 2013). On the other hand, all protein sequence alignments of TF genes were performed through stand-alone BLASTp version 2.2.18. Protein sequences of a template were compared with cassava protein library, called first BLASTp. Then, cassava proteins from the first BLASTp were subsequently aligned with protein library of a template genome, called

second BLASTp. The criteria of both alignments were set as: E-value≤1e-10, identity≥60%, and coverage≥80%. If cassava protein sequences passed these criteria with the reciprocal hit, they would be assigned function through *A. thaliana* genes annotation.

Verifying the binding possibility between predicted TF and target genes in cassava TRN via *cis*-regulatory element analysis

Cis-regulatory element analysis was performed to find binding possibility of TF in the 5' regulatory regions of target genes. The 1,500 bps from the translation start site (TLS) of each starch related genes were extracted from Phytozome database. TFBSs on upstream regions were analyzed with the Plant Promoter Analysis Navigator (PlantPAN) database (Chang et al., 2008). Regulation of TF to target gene was verified if its TF family from c*is*-regulatory element analysis and those of TF in the inferred TRN were the same. Co-expression profiles between TF genes and target genes were also demonstrated to consolidate the function of TF to target genes. Pearson's correlation coefficient (PCC) was applied to perform co-expression analysis by using gene expression data under cold stress of cassava apical shoots (An et al., 2012). The Affymetrix CEL files of microarray were downloaded from NCBI database (http://www.ncbi.nlm.nih.gov; experiment reference number GSE31073). Probe annotation was performed by stand-alone Two-way BLASTn between cassava probes in microarray chip and all transcript sequences in cassava genome. The criteria values of both alignments were set as: E-value≤1e-10, identity≥60%, and coverage≥80%.

Results

Characteristics of the TRN of starch metabolism in *A. thaliana*

The TRN of *A. thaliana* retrieved from AGRIS database (Davuluri et al., 2003) was used as a template to infer TRN of starch metabolism in cassava. The template network consists of 11,354 interactions from 67 TF genes and 8,130 target genes. Each TF can regulate one to 4,100 target genes. Thirty four of 67 TF genes can regulate only one target gene. Finally, 63 genes including 52 starch-related genes and 11 TF genes and their 79 interactions between TF and target gene were identified as TRN of starch metabolism in *A. thaliana* (Figure 2 (left)).

Inferred TRN of starch metabolism in cassava

The 63 genes in TRN of starch metabolism in *A. thaliana* were used as a template to identify orthologous genes in cassava by comparing all protein sequences of these genes with all cassava proteins in the genome via BLASTp version 2.2.18. Results show that only one TF gene (cassava4.1_017720m.g) was reciprocal hit and passed criteria. According to filtering TF-target gene pairs from previous step, two starch-related genes in cassava (cassava4.1_017170m.g and cassava4.1_017243m.g) were assigned to interact with a TF (Figure 2 (right)). This TF gene is transcription factor LONG HYPOCOTYL5 (HY5) classified in the basic-leucine zipper (bZIP) transcription factor family, while the other two target genes encode for ribulose-bisphosphate carboxylase, Rubisco (EC 4.1.1.39).

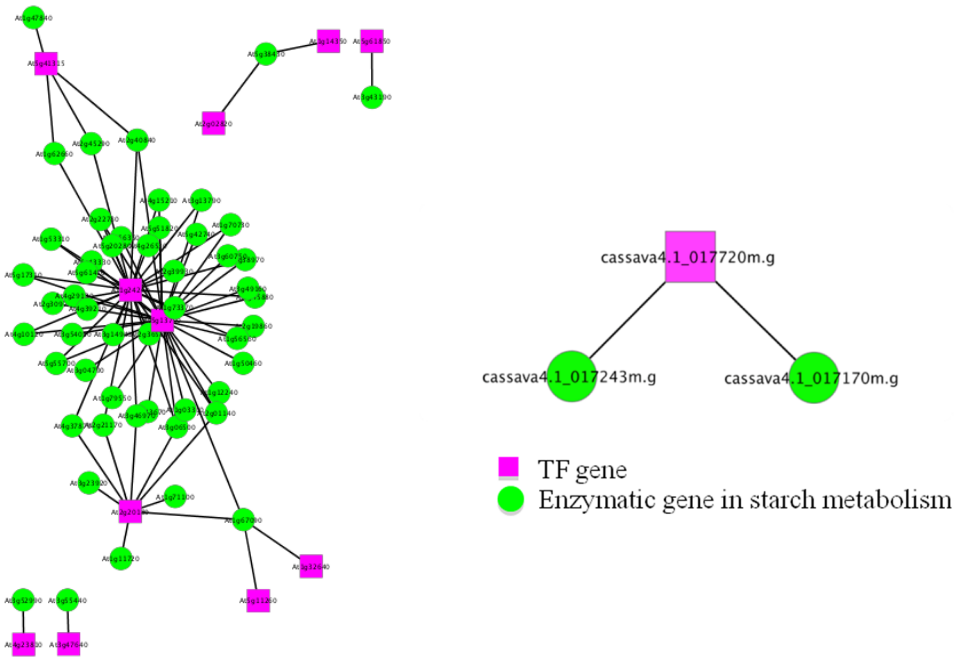

Figure 2: TRN of starch metabolism in *A. thaliana* (left) and cassava (right). Genes related to starch metabolism are represented as green circles, while TF genes are represented with a pink rectangle.

High quality TRN of starch metabolism in cassava

To verify the regulation of TF to starch-related genes, binding possibility of bZIP transcription factor (cassava4.1_017720m.g) on the promoter of two Rubisco genes (cassava4.1_017170m.g and cassava4.1_017243m.g) was analyzed. All possible *cis*-acting regulatory elements were identified on 1.5 kbps of 5' regulatory regions from the translation start sites (TLS) of the two target genes by using the plant transcription factor binding profiles in the PlantPAN database (Chang et al., 2008). The results demonstrated that the transcription factor HY5 classified in bZIP transcription factor family can bind at the promoter regions of the two ribulose-bisphosphate carboxylase genes as shown in pink rectangle (Figure 3). Co-expression patterns between two gene pairs, which were cassava4.1_017720m.g and cassava4.1_017243m.g (PCC= -0.59) and cassava4.1_017720m.g and cassava4.1_017170m.g (PCC= -0.14), were discovered from expression profile of cassava apical shoot under cold stress (Figure 4).

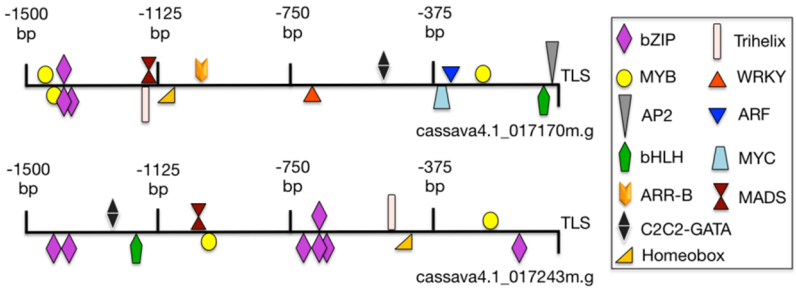

Figure 3: The *cis*-acting regulatory elements of possible TF families on the promoters of two Rubisco genes. TF families are represented as different shapes and colors and shown in the right box.

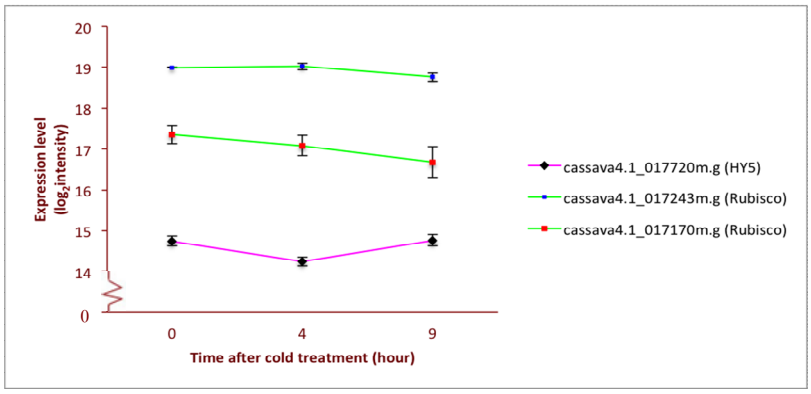

Figure 4: Expression profile of cassava4.1_017720m.g (HY5) and its targets, cassava4.1_017243m.g (Rubisco) and cassava4.1_017170m.g (Rubisco). All of them were annotated from probe named CUST_5455, CUST_9435 and CUST_9425 in An et al. (An et al., 2012) respectively via BLASTn.

Discussion

This work took the advantage of the available *A. thaliana* TRN to infer TRN of starch metabolism in cassava by using template-based method. Although, *A. thaliana* are not closely related to cassava based on evolution, it is a model plant providing experimental information of TRN, including both direct and indirect regulation between TF and target genes. With 11,354 interactions between 67 TF genes and 8,130 target genes, TF genes can regulate target genes, including enzymatic genes, TF genes, or other genes. Approximately, 50 percentages of TF genes can regulate only one target gene while the other group of TF genes can act as a global regulator by controlling a lot of target genes. In our study, genes involving in starch metabolism in *A. thaliana* were obtained that are the Calvin cycle, sucrose synthesis, and storage starch biosynthesis pathway. Eleven TF genes were assigned to be as regulators controlling 52 starch-related genes in starch metabolism of *A. thaliana*. Only one TF and one starch-related gene in starch metabolism of *A. thaliana* were shown the othologous in cassava. This might be caused from genetic variation of non-closely related organism and the limitation of experimental information of TF controlling starch-related gene in a template organism. Based on comparative analysis by using BLASTp, only one bZIP transcription factor (cassava4.1_017720m.g) gene was a potential regulator of two genes (cassava4.1_017170m.g and cassava4.1_017243m.g) encoding Rubisco. This enzyme catalyzes the first reaction in Calvin cycle (EC 4.1.1.39) in which D-ribulose-1,5-bisphosphate (RuBP) and carbon dioxide (CO_2) are converted into two molecules of 3-phosphoglycerate (3-PGA) (Gutteridge and Pierce 2006) which further used to produce fructose diphosphate. The fructose diphosphate is then used as substrate for glucose, sucrose, starch and other carbohydrates synthesis. So, Rubisco is the key enzyme for carbon fixation in plants and make plants to produce important primary and secondary compounds (Pandurangam et al., 2006). It was reported that sugar accumulation in leaves is related with the decreasing in expression of Rubisco gene (Sun et al., 2002). Decreasing in Rubisco gene expression under prolonged elevated CO_2 in plants was proposed to contribute to the down-regulation in photosynthesis.

The information between TF and transcription factor binding site (TFBS) is essential for drawing the clear picture of gene regulation. TFBSs of the predicted TF on these two target genes were verified by using PlantPAN database (Chang et al. 2008). PlantPAN is a database collecting the plant transcription factor binding profiles. The results were shown that this

predicted TF can bind on the upstream region of both target genes. Moreover, the regulation of TF to target genes was then consolidated via co-expression patterns between TF and target gene under cold stress of cassava apical shoots (An et al., 2012). Briefly, cassava plants were planted in plastic pots at 28°C under a 16 h light photoperiod for 3 months in the greenhouse. Next, plants with a uniform growth status were transferred to a chamber for cold treatment at 7°C under weak light. A microarray was used to measure transcriptome profiling in apical shoots of cassava for 0, 4 and 9 hours after cold treatment. Pearson's correlation coefficient (PCC) was applied to measure co-expression behavior between TF and two genes encoding Rubisco. An absolute PCC closed to one means a more significant regulatory relationship between TF and target genes. TF demonstrated a negative regulation to cassava4.1_017243m.g (PCC = -0.59) and negative regulation to cassava4.1_017170m.g (PCC = -0.14) as shown in Figure 4. Unfortunately, due to the limitation of transcriptome data in cassava, the regulation of TF to Rubisco genes was emerged in stress, affected by cold stress. Its regulation will be more clearly with the normal condition without stress for microarray experiment. However, this physical regulation between TF and these target genes have to be proved by experimental technique.

Conclusion

Cassava is one of the most important carbohydrate resources for human beings around the world, especially in Africa and Asia. Although cassava is one of the important starchy root crops, current understanding of starch biosynthesis pathway relies on biological investigation in the model plant, *A. thaliana*. TRN, revealing interactions between transcriptional factors and target genes involved in starch metabolism in cassava, was reconstructed here by applying the advantages of the template-based method. The known TRN of *A. thaliana* in the AGRIS database was used as a template for cassava TRN inference. Seventy-nine of 11,354 well-known interactions between 52 starch-related genes and 11 transcription factors were identified in *A. thaliana*. The comparative genomic approach was then used to identify orthologous genes in cassava. The results proposed that two cassava genes, cassava4.1_017170m.g and cassava4.1_017243m.g, encoding ribulose-bisphosphate carboxylase (Rubisco), were controlled by putative transcription factor LONG HYPOCOTYL5 (HY5) in the basic-leucine zipper (bZIP) transcription factor family. This TF can bind on the upstream region of the two Rubisco genes through *cis*-regulatory element analysis. The regulation function of TF on the two target genes was then consolidated by gene co-expression analysis between TF and Rubisco genes under cold stress of cassava apical shoots. However, the physical regulation between TF and these target genes have to be proved by experimental work.

Acknowledgements

This work is supported by Thailand research cluster (NSTDA, TRF, ARDA, STI, HSRI and NRCT) with grant ID: P-12-00743. We would like to thank the National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand for supporting postgraduate scholarship to Bandit Khampoosa.

References

An D, Yang J, Zhang P (2012) Transcriptome profiling of low temperature-treated cassava apical shoots showed dynamic responses of tropical plant to cold stress. BMC Genomics. 13:64.

Bumee S, Ingkasuwan P, Kalapanulak S, Meechai A, Cheevadhanarak S, Saithong T (2013) Transcriptional Regulatory Network of Arabidopsis Starch Metabolism under Extensive Light Condition: A Potential Model of Transcription-modulated Starch Metabolism in Roots of Starchy Crops. Procedia Computer Science. 23:113-121.

Chang WC, Lee TY, Huang HD, Huang HY, Pan RL (2008) PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. BMC Genomics. 9:561.

Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics. 4:25.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2011) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research.

Gutteridge S, Pierce J (2006) A unified theory for the basis of the limitations of the primary reaction of photosynthetic CO(2) fixation: was Dr. Pangloss right? Proc Natl Acad Sci U S A. 103:7203-7204.

Huang Y, Li H, Hu H, Yan X, Waterman MS, Huang H, Zhou XJ (2007) Systematic discovery of functional modules and context-specific functional annotation of human genome. Bioinformatics. 23:i222-i229.

Ingkasuwan P, Netrphan S, Prasitwattanaseree S, Tanticharoen M, Bhumiratana S, Meechai A, Chaijaruwanich J, Takahashi H, Cheevadhanarak S (2012) Inferring transcriptional gene regulation network of starch metabolism in Arabidopsis thaliana leaves using graphical Gaussian model. BMC Syst Biol. 6:100.

Janky R, Helden J, Babu MM (2009) Investigating transcriptional regulation: from analysis of complex networks to discovery of cis-regulatory elements. Methods 48:277-286.

Li K, Zhu W, Zeng K, Zhang Z, Ye J, Ou W, Rehman S, Heuer B, Chen S (2010) Proteome characterization of cassava (Manihot esculenta Crantz) somatic embryos, plantlets and tuberous roots. Proteome Science. 8:10.

Liu J, Xu X, Stormo GD (2008) The cis-regulatory map of Shewanella genomes. Nucleic Acids Res. 36:5376-5390.

Mao L, Van Hemert J, Dash S, Dickerson J (2009) Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics. 10:346.

Pandurangam V, Sharma-Natu P, Sreekanth B, Ghildiyal MC (2006) Photosynthetic acclimation to elevated CO2 in relation to Rubisco gene expression in three C3 species. Indian J Exp Biol. 44:408-415.

Saithong T, Rongsirikul O, Kalapanulak S, Chiewchankaset P, Siriwat W, Netrphan S, Suksangpanomrung M, Meechai A, Cheevadhanarak S (2013) Starch biosynthesis in cassava: a genome-based pathway reconstruction and its exploitation in data integration. BMC Syst Biol. 7:1-17.

Sun J, Gibson KM, Kiirats O, Okita TW, Edwards GE (2002) Interactions of nitrate and CO2 enrichment on growth, carbohydrates, and rubisco in Arabidopsis starch mutants. Significance of starch and hexose. Plant Physiol. 130:1573-1583.

Wanatsanan S, Kalapanulak S, Suksangpanomrung M, Netrphan S, Meechai A, Saithong T (2012) Transcriptomic data integration inferring the dominance of starch biosynthesis in carbon utilization of developing cassava roots. Procedia CS. 96-106.

Yang J, An D, Zhang P (2011) Expression profiling of cassava storage roots reveals an active process of glycolysis/gluconeogenesis. J Integr Plant Biol. 53:193-211.