"COMPARATIVE AND FUNCTIONAL GENOMICS ANALYSIS OF STARCH BIOSYNTHESIS PATHWAYS IN CASSAVA"

By

POOJA HARSHAN

(2012-09-102)

Thesis

Submitted in partial fulfilment of the

Requirement for the degree of

MASTER OF SCIENCE (INTEGRATED) IN BIOTECHNOLOGY

Faculty of Agriculture
Kerala Agricultural University, Thrissur

B. Sc. - M. Sc. (INTEGRATED) BIOTECHNOLOGY

DEPARTMENT OF PLANT BIOTECHNOLOGY

COLLEGE OF AGRICULTURE

VELLAYANI, THIRUVANANTHAPURAM - 695 522

KERALA, INDIA

2017

DECLARATION

I hereby declare that this thesis entitled "**Comparative and Functional Genomics Analysis of Starch Biosynthesis Pathways in Cassava**" is a bonafide record of research work done by me during the course of research and that the thesis has not previously formed the basis for the award of any degree, diploma, associate ship, fellowship or other similar title, of any other university or society.

Place: Vellayani
Date: 31/10/2017

POOJA HARSHAN
(2012-09-102)

CERTIFICATE

Certified that this thesis entitled **"Comparative and Functional Genomics Analysis of Starch Biosynthesis Pathways in Cassava"** is a record of research work done by **Ms. Pooja Harshan (2012-09-102)** under my guidance and supervision and this is not previously formed the basis for the award of any degree, diploma, fellowship or associate ship to her.

Place: ICAR-CTCRI, Sreekariyam
Date: 31/10/2017

Dr. J. Sreekumar
(Chairman, Advisory Committee)
Principal scientist
Section of Extension & Social Sciences
ICAR-CTCRI, Sreekariyam,
Thiruvananthapuram, 695017

Produce tuber reduce hunger

भाकृ अनुप-केन्द्रीय कन्द फसल अनुसंधान संस्थान
(भारतीय कृषि अनुसंधान परिषद)
श्रीकार्यम, तिरुवनन्तपुरम -६९५०१७, केरल,भारत

ICAR-CENTRAL TUBER CROPS RESEARCH INSTITUTE
(Indian Council of Agricultural Research)
Sreekariyam, Thiruvananthapuram-695 017, Kerala, India

ISO 9001:2008

CERTIFICATE

Certified that this thesis entitled **"Comparative and functional genomic analysis of starch biosynthesis pathway in cassava (*Manihot esculenta C.*)"** is a record of research work done by Ms. Pooja Harshan (2012-09-102) under my guidance and supervision and that is not previously formed the basis for the award of any degree, diploma, fellowship or associateship to her.

Sreekariyam
Date: 26.08.2017

Dr. J. Sreekumar
(Chairman, Advisory Committee)
Senior Scientist (Agrl. Statistics)

4

Phone : 91-0471-2598551 to 2598554
Director (Per.) : 91-0471-2598431
(Res.) : 91-0471 - 2441957

RIGHT TO INFORMATION

Fax : 91-0471-2590063
E-Mail : ctcritvm@yahoo.com
ctcritvm@gmail.com

CERTIFICATE

We, the undersigned members of the advisory committee of Ms. Pooja Harshan, a candidate for the degree of B.Sc. M.Sc. (Integrated) Biotechnology, agree that the thesis entitled "**Comparative and Functional Genomics Analysis of Starch Biosynthesis Pathways in Cassava**" may be submitted by Ms. Pooja Harshan, in partial fulfilment of the requirement for the degree.

3/11/17

Dr. J. Sreekumar
(Chairman, Advisory Committee)
Principal scientist
Section of Extension & Social Sciences
ICAR-CTCRI, Sreekariyam
Thiruvananthapuram

Dr. Swapna Alex
(Member, Advisory Committee)
Professor & Head
Department of Plant Biotechnology
College of Agriculture, Vellayani
Thiruvananthapuram

Dr. C. Mohan
(Member, Advisory Committee)
Principal Scientist
Division of Crop Improvement
ICAR-CTCRI, Sreekariyam
Thiruvananthapuram

03/11/17

Dr. A. N. Jyothi
(Member, Advisory Committee)
Principal scientist
Division of Crop Utilization
ICAR-CTCRI,Sreekariyam
Thiruvananthapuram

EXTERNAL EXAMINER

3/11/17

Dr. Rajesh M. K.
Principal scientist
Agricultural Biotechnology
Division of Crop Improvement
ICAR-CPCRI, Kasaragod

5

ACKNOWLEDGEMENT

It is with my heartfelt feelings; I would like to express my sincere gratitude to my beloved advisor Dr. J. Sreekumar for his continuous support, patience and motivation. In addition to his support for completing the project, his advice was valuable for me in every aspect.

My special thanks to former Director of ICAR-CTCRI Dr. James George and present Director Dr. Archana Mukherjee for allowing me to do my project work and also for their support.

I take immense pleasure to express my deep sense of gratitude to Prof. B. R. Reghunath, Dr. R. Rajmohan, Dr. Swapna Alex, Dr. C. Mohan and Dr. A. N. Jyothi not only for their insightful comments and encouragement, but also for their valuable counseling and constructive suggestions that were much helpful throughout my research progress. My sincere thanks also goes to Dr. K. B. Soni, our course Director for her encouragement and help rendered during the course of the study. I would also like to put on record my sincere thanks and gratitude to Dr. Sheela Immanuel, Head, Section of Extension and Social Sciences, ICAR-CTCRI for permitting me and extending all facilities to complete my work.

My sincere gratitude towards Mr. Ambu Vijayan, who helped me with the technical aspects of my work including the pathway reconstruction and gene prioritization using various bioinformatics tools and software. My heartfelt thanks are also due to Dr. Visalakshi Chandra, Mr. Vishnu S. Nath, Mr. Nidhin A.N, Ms Vinutha, Mrs. Sreelatha and Mrs. Sujina. My special thanks to my friends Aishwarya B. S., Muhammed Sadiq P. A., Reshmi B. S. and Rekha

Jayaraj for helping me in completing this project work. My acknowledgement would be lacking if I don't mention my gratitude to my beloved friends, Akshara George, Pooja A. P., Pallavi Thakur, Pooja Das, Chanchal Maria Joseph, Gopika, Neenu, Adarsh Prathap and all other classmates for their invaluable care, constant support, motivation and selfless help.

I wish to express my deep gratitude to all non-teaching staff members of ICAR – CTCRI, my seniors, juniors and teachers in college for their timely help.

I acknowledge the favor of numerous persons who, though not been individually mentioned here, who have all directly or indirectly contributed to this work. I owe a great deal to my dear mother Sunita and uncle Sudhir who stood with me in all my happiness, difficulties and sorrow.

I'm grateful to my dear parents and my brothers, for their love, blessing, prayers and encouragement.

Finally, I humbly thank the Almighty for showering his blessings and bestowing the wisdom, perseverance and physical ability to accomplish this work.

POOJA HARSHAN

TABLE OF CONTENTS

Sl. No.	Chapters	Page No.
1	INTRODUCTION	1-3
2	REVIEW OF LITERATURE	4-19
3	MATERIALS AND METHODS	20-28
4	RESULTS	29-46
5	DISCUSSION	47-51
6	SUMMARY	52-54
7	REFERENCES	55-79
8	APPENDICES	80
9	ABSTRACT	82

8

LIST OF TABLES

Table No.	Title	Page No.
1	Genes involved in carbon-dioxide fixation in cassava	30
2	Genes involved in sucrose synthesis I in cassava	31
3	Genes involved in sucrose synthesis II in cassava	32
4	Genes involved in starch synthesis in cassava	33
5	Function assignment to the predicted sequences using E2P2 tool	34
6	Protein motifs identified linked to GO terms using GOMO tool	36
7	Ranking of motifs from alpha 1,4 glucan branching enzyme using TOMTOM tool	37
8	Promoter region prediction in predicted alpha 1,4 glucan branching enzyme using PROSCAN tool	39
9	Transcription factor binding sites predicted using JASPAR for putative alpha 1, 4 glucan branching enzyme	40
10	Transcription factor binding sites predicted using PlantTFDB for putative alpha 1, 4 glucan branching enzyme	41
11	Candidate genes involved in starch biosynthesis in cassava predicted through ToppGene suite	45
12	Differentially expressed genes identified from microarray gene expression analysis	46

LIST OF FIGURES

Figure No.	Title	Between pages
1	Workflow for the starch biosynthesis pathway reconstruction in cassava	22-23
2	Motifs identified using MEME tool	39-40
3	Protein-protein interaction network of predicted Starch synthase created using STITCH	39-40
4	Protein-protein interaction network of predicted UDP glycosyl transferase created using STITCH	39-40
5	Transcription factor binding site prediction using Alggen-PROMO in the predicted alpha 1,4 glucan branching enzyme	41-42
6	Consolidated pathway of starch biosynthesis constructed using Cell Illustrator	42-43
7	Consolidated pathway of starch biosynthesis viewed on VANTED tool	45-46
8	Heat map of the genes after normalization generated from GeneXplain platform	45-46
9	HCL clustering of genes after normalization generated from MeV software	45-46
10	Heat map of the significant genes generated from MeV software	45-46
11	SAM graph of the significant genes from MeV software	46-47
12	Volcano plot of the significant genes from MeV software	46-47
13	The interactive pathway of reconstructed starch biosynthesis pathway in cassava along with the integration of differentially expressed genes viewed on VANTED platform	46-47

LIST OF APPENDICES

SL. No.	Title	Appendix No.
1	E2P2(Ensemble Enzyme Prediction Pipeline version 3.0) SOFTWARE INSTALLATION	I
2	E2P2(Ensemble Enzyme Prediction Pipeline version 3.0) SOFTWARE OPERATION	I

LIST OF ABBREVATIONS

ADP	Adenosine diphosphate
AGPase	ADP-glucose Pyrophosphorylase
APL	ADP-glucose Pyrophosphorylase Large Subunit
APS	ADP-glucose Pyrophosphorylase Small Subunit
BIMAS	Bioinformatics and Molecular Analysis Section
BLAST	Basic Local Alignment Search Tool
cDNA	Complementary Deoxyribonucleic Acid
CIBEX	Center for Information Biotechnology Gene Expression Database
DBE	Debranching Enzyme
DEGs	Differentially Expressed Genes
DR	Developing storage Root
DW	Dry Weight
E2P2	Ensemble Enzyme Prediction Pipeline
EBI	The European Bioinformatics Institute
EF	Enzyme Function
ESTs	Expressed Sequence Tags
FASTA	FAST Alignment
FR	Fibrous Root

GBSS	Granule-Bound Starch Synthase
GEO	Gene Expression Omnibus
GH	Glycoside Hydrolase
GO	Gene Ontology
GOMO	Gene Ontology for Motifs
GWD	Glucan Water Dikinase
HCL	Hierarchial Clustering
ISA	Isoamylase
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAST	Motif Alignment and Search Tool
MEME	Multiple Em for Motif Elicitation
MeV	Multi Experimental Viewer
MR	Mature storage Root
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NIG	National Institute of Genetics
PGA	Phosphoglycerate
P_i	Inorganic phosphate
PlantTFDB	Plant Transcription Factor Database
PMN	Plant Metabolic Network

PROSCAN	Promoter Scan
RNA-seq	RNA sequencing
SAM	Significance Analysis of Microarray
SBE	Starch Branching Enzyme
SNR	Sound to Noise Ratio
SS	Soluble Starch Synthase
TAIR	The Arabidopsis Information Resource
VANTED	Visualization and Analysis of Networks containing Experimental Data
WSG	Water Soluble Glucan

14

INTRODUCTION

1. INTRODUCTION

Cassava is one among the most important tropical crops grown on a global scale next to rice and maize (FAO, 2008). Over 700 million people in Africa (51%), Asia (29%) and South America (20%) consume it (FAOSTAT, 2017). The most attractive feature that makes cassava different from other starch crops is the enhanced production per hectare in the form of carbohydrate reservoir and energy demand (Montagnac *et al.*, 2009). Cassava is unique in various aspects such as the capacity to produce competitive yields under poor soils, salinity, low pH or acidic conditions, drought, high air temperatures and evapotranspiration and pest attack (El-Sharkawy, 2006; Burns *et al.*, 2010; FAO, 2013).

Cassava is also an important source of commercial starch that is used by various industries for the production of value-added products (Da *et al.*, 2013). The starch produced by cassava is of elite quality that makes it a potential candidate by various industries like paper, textiles, food processing, pharmaceuticals and for ethanol production (Ceballos *et al.*, 2004). Cassava breeding programme was initiated with the main objective to develop novel cassava lines with superior starch content (Kunkeaw *et al.*, 2011) with variation in the starch granule morphology and functional characteristics of starch (Ceballos *et al.*, 2012) and develop herbicide tolerant varieties (Taylor *et al.*, 2004).

The starch biosynthesis involves the production of amylose and amylopectin molecules catalyzed by an enzyme cascade. Amylose synthesis is carried out with the aid of *GBSSI* (Granule bound starch synthase) enzyme (Buleon *et al.*, 1998). The production of amylose free starch involves the silencing or knocking out the *GBSSI* gene thereby producing *waxy* cassava (Jobling, 2004). However, the synthesis of amylopectin moiety of the starch is associated with another chain of enzymes namely four soluble starch synthases like *SSI*, *SSII*, *SSIII* and *SSIV*, two starch branching enzymes such as *SBEI* and *SBEII,* the Glucan Water Dikinase (GWD) and several debranching enzymes and kinases

(Brummel *et al.*, 2015). The starch synthases and starch branching enzymes add glucose moieties to the main chain and assist in the cleavage as well as introducing branching points in the amylopectin molecules (Martin and Smith, 1995). The expression of different isoforms of starch synthases is highly dependent on the prevailing environment conditions as well as the genotype of the crop. GWD is the rate limiting enzyme in starch degradation and controls the amount of starch accumulation in the plant. Plants lacking the GWD enzyme stores invariably large amounts of starch (Skeffington *et al.*, 2014).

High-throughput technologies revolutionised the investigations on the system biology of plants. Several system biology approaches such as genomics, proteomics and transcriptomics have been employed in understanding the mechanism underlying starch biosynthesis and disentangle the complex regulatory processes in the crop (Sakurai *et al.*, 2007; Yang *et al.*, 2011). Using comparative genomics approach and cassava genome data, the investigations on starch biosynthesis was very well studied at the pathway level (Goodstein *et al.*, 2011). The starch synthesis pathway reconstructions carried out by Rongsirikul *et al.* (2010) and Saithong *et al.* (2013) further added more insights into the mechanism of starch synthesis in cassava. Transcriptome data integration enabled the identification of differentially expressed genes in cassava during different growth stages of roots. However, both the reconstruction protocols lacked the molecular mechanism underlying starch synthesis which includes the key genes, regulatory elements involved and also the candidate genes participating in the pathway.

A new strategy was employed in the study for the identification of candidate genes named Gene Prioritisation which helps in screening the key genes involved in the starch synthesis from a large subset of genes. Genotype-phenotype relationship serves as the basis for any plant breeding programmes. The associations between the plant characteristic traits and biological processes which are either over-represented or enriched help in narrowing down the number of potential candidate genes for that particular trait. Also the functional regulatory

element analyses were carried out to provide a better understanding of the pathway at the transcriptional level along with the incorporation of microarray gene-expression data. Hence, the present study aimed at developing an integrated pathway composing of the predicted genes, regulatory elements and also the gene-expression data. The reconstructed pathway could be used by the plant breeders to identify the differentially expressed genes in different growth stages of cassava and understand the starch metabolism for the development of elite novel cassava lines with higher starch content.

REVIEW OF LITERATURE

2. REVIEW OF LITERATURE

2.1 ROOTS AND TUBER CROPS

Starchy tubers and root crops serve second position to cereals as potential source of carbohydrates. They provide a considerable source of global supply chain, animal feed and other processed materials. There are a wide variety of tuber crops *viz.,* cassava, potatoes, sweet potatoes, yams etc. Potatoes and yams are modified tubers, taro and cocoyam are modified corms, cassava and sweet potatoes are storage roots and canna and arrowroots are edible rhizomes. These crops are of high national importance in terms of global annual production which accounts to 836 million tonnes (FAOSTAT, 2013). Asia and Africa produces 43 and 33 percent respectively of the globally available roots and tubers (FAOSTAT, 2013). However, the protein content of roots and tubers is considerably low in the range of 1 to 2 % on a dry weight basis (Burlingame *et al.*, 2009). Roots and tubers are a rich reservoir of dietary fibre (FAO, 1990). Among the root crops, cassava, sweet potatoes and potatoes are produced on a remarkably large scale that constitutes 90% of the global production of roots and tuber crops (FAOSTAT, 2013).

2.2 CASSAVA

Cassava (*Manihot esculenta* Crantz.) is a member of the family Euphorbiaceae and the morphological appearance is a shrub. Euphorbiaceae family comprises of plants growing in diverse habitats like trees, shrubs and grasses. The economically important plants in the Euphorbiaceae family include rubber, castor-bean and cassava. Cassava is cultivated mainly in tropics (Anderson *et al.*, 2004) and is the staple food crop to over one billion people in around 105 countries (Chetty *et al.*, 2013). Cassava is ranked as the fourth important food crop on a global scale after rice, sugarcane and maize (Baguma, 2004). Apart from the prospective use of cassava in being a food crop, it finds applications in various industries including cosmetics, bio-fuel and biopolymer

industries (Balat and Balat 2009). Cassava is highly competitive when compared with other starchy crops in its immense ability to accumulate starch in its roots, being drought tolerant and is also deficit of soil nutrients (Okogbenin *et al.*, 2013). Being a drought tolerant crop, cassava can be grown in a broad range of agro-ecologies offering a comparatively cheap cost vegetative propagation during different harvest cycles (Uchechukwu-Agua *et al.*, 2015). However, relatively higher levels of cyanogenic compounds, low levels of proteins and micronutrients and postharvest physiological deterioration are the major constraints to the economic development of cassava (Vanderschuren *et al.*, 2014). Cassava plant is also susceptible to bacterial (Verdier *et al.*, 1994) and viral diseases (Patil and Fauquet 2009) that are transmitted by insects. To extend the market value of cassava, it is normally processed into cassava starch leading to the development of value-added products by different sectors of food industries. Various cassava breeding programs have been implemented with the aim of developing new cultivars with increased starch content and a variation in the functional properties of starch.

2.3 STARCH BIOSYNTHESIS

Starch is a major form of stored carbohydrate in plants. Starch is an assembled form of storage metabolite in plants composed of α 1, 4 linear glucan or amylose which is least branched (Zeeman *et al.*, 2010) and α 1, 6 branched glucan or amylopectin (Smith, 2001). The degree of polymerisation (dp) of amylose molecules are species dependent and ranges between 800-4500 (Alexander, 1995). In contrast to amylose, amylopectin molecules have degree of polymerisation in the range of 10^5 to 10^7 (Myers *et al.*, 2000). The amylose content in starch is in the range of 17-30% and the amylopectin composition is around 70-83%. Gerard *et al.* (2002) has developed certain mutant plants with a waxy endosperm having a lowered or no amylose content.

2.3.1 Cassava starch

Cassava starch resembles other sources of starch in its gelatinisation and retrogradation properties. Gelatinisation property is characterised by the solvent type and also the starch/solvent proportion (Jimenez *et al.*, 2012). Cassava is a rich source of vitamin C and minerals and contains at an approximate of 800 g of starch per Kg of dry weight of the tuber (Benesi *et al.*, 2004). The nutrient composition of cassava is dependent on various factors like processing conditions, cultivars and harvesting periods. The starch quality of cassava is based on the amylose content in the starch (Srichuwong *et al.*, 2007). High-amylose starches are more hydrolysis resistant when compared with starches having high amylopectin content. The main by-product of cassava starch which is the cassava pulp that is a rich source of starch (400-600g/kg DW) finds application in the production of value-added products with better bio-refinery concepts (Patle and Lal 2008).

2.3.2 Starch biosynthesis in cereals

The precursor of starch biosynthesis, ADP- glucose is synthesized in the chloroplast/plastids in all the plants including cereals. However, in the developing endosperms of the cereals, the synthesis of ADP-glucose was found to take place in both the plastid as well as the cytosol. The import of ADP-glucose from the cytosol to the plastid is facilitated with the help of an antiporter named ADP-glucose transporter which is present exclusively in grasses. Hence, the grass endosperm is characterised by the presence of both the isoforms of the AGPase enzyme *viz.,* plastidial and cytosolic.

Giroux and Hannah (1994) reported that the AGPase enzyme was purely cytosolic in the case of cereal endosperm. Studies of the *brittle* (bt_1) mutants of maize by Shannon *et al.*, (1996) showed that the mutants were able to accumulate ADPglucose 13 times more than the normal amount of ADPglucose without the formation of normal starch. Thorbjornsen *et al.* (1996) identified two different transcripts that were encoded by a single small unit of the AGPase enzyme

shedding light on the existence of an independent cytosolic AGPase enzyme in the endosperm of barley.

2.3.3 Starch biosynthesis in non-cereals

In plants other than cereals, the starch is produced from the end products of photosynthesis which is later translocated from the source organs like leaves to the sink organs like stems and roots. The translocation of glucose 6-phosphate and ATP, which are the suitable substrates for starch biosynthesis take place from the cytosol to plastid via glucose 6-phosphate/phosphate transporter and the ATP/ADP transporter (Neuhaus and Wagner 2000). Tetlow *et al.* (1996) reported that glucose 6-phosphate/phosphate transporter is capable in transporting glucose 1-phosphate into the plastids. Hatzfeld and Stitt (1990) conducted experiments on the starch biogenesis in wheat and maize grains using ^{13}C NMR spectroscopy and concluded that the precursor of starch biosynthesis in amyloplasts is a 6 carbon molecule rather than a 3 carbon Triose-phosphate as in the case plastids. Frehner *et al.* (1990) provided strong experimental evidences supporting the precursor in the amyloplasts as a 6 carbon moiety through investigations focused on analysing the enzyme activities of the amyloplasts.

Rolletschek *et al.*, (2002) conducted metabolic flux analysis in heterotrophic organs and suggested that the reaction catalysed by AGPase gene is not the rate limiting step in starch synthesis. He obtained a flux control coefficient for AGPase as 0.08 which made him to draw the above conclusion. Tiessen *et al.* (2003) studied on the post translational redox modifications and 3PGA/Pi balance in tubers of potato shedding light on the regulation of starch synthesis in potato. Tjaden *et al.* (1998) reported the presence of an ATP/ADP translocator in the envelope membranes of amyloplasts. The evidence on the potential function of ATP/ADP translocators in starch synthesis came from the work conducted by Geigenberger *et al.* (2001) who reported that potato tubers exhibiting a decreased plastidial ATP/ADP translocator has got a reduced starch content.

Tang and Sturm (1999) studied the importance of *SuSy* in the starch-sucrose conversion and concluded that potato tubers and carrot roots with lower expression of *SuSy* enzyme lead to the accumulation of reduced amount of starch in the plants. The amount of *in vitro* ADPG which have a potential to produce *SuSy* activity is maximum for starch containing storage organs when compared with that of potato tubers (Baroja-Fernandez *et al.*, 2009) and 3-4 fold change in the production of ADP was observed when compared with the developing barley and maize endosperms (Li *et al.*, 2013).

2.4 ENZYMES INVOLVED IN STARCH BIOSYNTHESIS

2.4.1 ADP-glucose pyrophosphorylase (2.7.7.27)

ADP-glucose pyrophosphorylase is the enzyme involved in the rate limiting step of starch biosynthesis in plants giving the enzyme a special attention among the biochemists (Tuncel and Okita, 2013). In plants, ADP-glucose pyrophosphorylase (AGPase) exists as a hetrotetramer composed of four subunits *viz.*, 2 large and 2 small (Hannah, 1997). Large subunits were subjected to more duplication events when compared with the small subunits (Georgelis *et al.*, 2008). Duplication events in these large subunits resulted in the creation of five clusters of large subunits that differ in their expression pattern in different tissues (Akihiro *et al.*, 2005). Crevillen *et al.* (2003) studied the biochemistry and also the functional properties of four different Arabidopsis AGPases consisting of the four diverse large subunits and a single functional small subunit in Arabidopsis. Tiessen *et al.* (2012) reported that large subunits have a defective catalytic property in potato, arabidopsis and barley.

Glucose-1-phosphate and ATP act as precursors for the synthesis of starch and is catalyzed by AGPase enzyme (Espada, 1962). AGPase is an allosteric enzyme which has got both the positive as well as negative regulation. The positive regulator of AGPase is 3-Phosphoglycerate (3-PGA) whereas the negative regulator is inorganic orthophosphate (P_i) (Hannah, 1997). The isoforms of AGPase that are seed and leaf specific have also been identified (Beckles,

2001). In cereals, AGPase that is present in the endosperm is devoid of transit peptides leading to the presence of both endosperms localized and chloroplast localized form of both the large and small subunits of AGPase enzyme.

Several studies reported that the over-expression of AGPase enzyme in plants lead to an increased strength of sink organs. Smidansky *et al.* (2007) reported that increased expression of a truncated form of large subunit of AGPase in maize under an endosperm-specific promoter led to increased seed weight and plant biomass. However, a significant change in the phenotype was not observed under field conditions (Meyer *et al.*, 2007). Rosti *et al.* (2007) studied the impact of a transposon derived AGPase silencing on the plant yield in rice. They observed a significant reduction in the amount of starch being produced in the plant whereas little or no change was seen for the plant growth or productivity. Another research conducted by Slewinski *et al.* (2008) showed that a similar sort of mutation in maize named the agps-m1 resulted in the lack of leaf starch due to the transposon insertion in the small subunit of leaf AGPase. Lee *et al.* (2009) reported an increased fresh weight and starch production in lettuce eight weeks after germination due to the over-expression of potato AGPase large subunit, *upreg1*. Similar study by Gibson *et al.* (2011) noticed a significant increase in the crop yield in rice when compared with the control on over-expression of leaf specific *upreg1*. Tsubone *et al.* (1997) pointed out that AGPase is crucial for determining sink activity and starch production in the roots of sweet potato (*Ipomoea batatas* (L.) Lam.).

Studies have revealed the importance of AGPase enzyme in regulating multiple biological processes such as growth, development, environmental adaptation and host-pathogen interactions in different plant species. In *Arabidopsis thaliana*, *AtAGPase* has a significant role in the enhancement of growth rate in young plantlets and also the root development (Mugford *et al.*, 2014). Kwak *et al.* (2007) studied the role of *NtAGPase* in petal expansion and growth in *Nicotiana tabacum*. A harmonious expression of *APS* along with the presence of starch in cotton (*Gossypium hirsutum*) fibre revealed the importance

of *GhAGPase* in the fibre development in cotton (Taliercio, 2011). Petreikov *et al.* (2006) reported that in *Solanum lycopersicon*, temporally expanded expression of *APL1* gene led to increased accumulation of starch and soluble sugars. The expression of AGPase genes in plants were found to be dependent on various factors such as stress, cold, drought, salinity and pathogen interactions (Yin *et al.*, 2010; Gamm *et al.*, 2011; Wiberley-Bradford *et al.*, 2014; Saripalli and Gupta, 2015). Low storage temperature stimulated the expression of AGPase gene in *S. tuberosum* (Wiberley-Bradford *et al.*, 2014). In *S. lycopersicon, APL1* gene was up-regulated during saline stress (Yin *et al.*, 2010). Drought stress led to a decrease in the crop yield (Thitisaksakul *et al.*, 2012; Majoul-Haddad *et al.*, 2013). However, downy mildew infection in grapevine leaves led to an increased level of AGPase enzyme and also abnormal starch accumulation (Gamm *et al.*, 2011).

Also the identification of key AGPase genes which have a potential role in the fruit development as well as in combating biotic and abiotic responses was done in many of the fruit crops including banana (Backiyarani *et al.*, 2015; Hu *et al.*, 2016). Genome-wide analysis has identified a variable number of *APL* and *APS* in a wide variety of plants belonging to higher strata. Four *APLs* and two *APS* have been identified in *Arabidopsis thaliana* (Crevillen *et al.*, 2005), Four *APLs* and three *APS* in *Oryza sativa* (Lu and Park, 2012; Tang *et al.*, 2016), Four *APLs* and two *APS* in *Zea mays*, Four *APLs* and two *APS* in *Ipomoea batatas* L. (Zhou *et al.*, 2016) and three APLs and one APS in *Actinidia deliciosa* (Nardozza *et al.*, 2013).

2.4.2 Starch synthases (2.4.1.21)

Starch synthases are the major enzymes involved in the extension of the glucan chains by the transfer of glucose residues from ADP-glucose to the non-reducing end of the growing glucan chains. Based on the presence of identical amino acid sequences, there are multiple isoforms of starch synthases in higher plants (Ball and Morell, 2003). Most of the starch synthases share a common

structure with a glass domain, a transit peptide and different motifs (Vrinten and Nakamura, 2000). Starch synthases are classified into three types in the plastids namely granule-bound, soluble and both granule-bound as well as soluble. Granule bound starch synthases I (*GBSSI*) exist as two isoforms in the plant such as *GBSSIa* and *GBSSIb* (Edwards *et al*., 2002). In plants, starch synthases are encoded by five conserved classes of genes (Leterrier *et al*., 2008). Denyer *et al.* (1996) reported that the *GBSSI* activity in plants is enhanced by MOS molecules. Seung *et al.* (2015) identified a plastidial protein named *PTST* which has got a significant role in amylose synthesis and also signalling *GBSSI* to starch granules in *Arabidopsis thaliana*. Arabidopsis *ptst* mutants phenotypically resembled *GBSSI* mutants. Four different isoforms of starch synthases like *SSI, SSII, SSIII* and *GBSSI* exist in potato. *SSI* has got an important role in starch biosynthesis and has no multiple isoforms like other starch synthases (Jeon *et al*., 2010). In potato, *SSI* was seen to be present in the leaves and has got significant role in the production of transitory starch (Kossmann *et al*., 1999). The synthesis of short strands of amylopectin molecules are catalysed by *SSI* (Delvalle *et al*., 2005). Edwards *et al.* (1999) detected a considerable change in the starch granule morphology when the *SSIII* enzyme was down-regulated. Differences in the structure of starch varies with species and is contributed by the arrangement in the isoforms of starch synthases between two compartments namely the stroma and the starch granules inside the plastids (Ball and Morell, 2003). Zhang *et al.* (2008) investigated on the functions of *SSII* and *SSIII* and was found to be similar.

Genetic modifications in the starch synthases lead to change in the granule structure, productivity and physical properties of starch. Craig *et al.* (1998) reported that an inhibition on the activity of *SSII* in rug5 mutants of pea resulted in an increase in the number of very long chains of amylopectin molecules, a substantial decrease in short chains and a complete loss of intermediate chains. When *SSII* has a major role in potato in the absence of *SSIII*, it was found that the starch phosphorylation as well as peak viscosity was increased to a larger extent.

Mutations in the *GBSSIa* genes produce waxy crops like waxy corn (Nelson and Pan, 1995) and glutinous rice (Olsen and Purugannan, 2002). In barley (*Hordeum vulgare)*, only *GBSSIa* is expressed in the endosperm (Radchuk *et al.*, 2009) whereas *GBSSIb* is expressed in the pericarp and the embryo (Sreenivasulu *et al.*, 2006). Hebelstrup *et al.* (2010) reported a transit peptide in barley which was formed out of first 69 N-terminal amino acids of *GBSSIa* that is capable of directing a transgenic GFP to the amyloplast of barley endosperm cells.

2.4.3 Starch branching enzyme (2.4.1.18)

Branching enzymes are a class of enzymes which cleave the α (1, 4) linkage of donor glucan chain and transfering the cleaved fragment into the acceptor fragment introducing α (1, 6) branch point in the acceptor chain. Acceptors are either part of the original glucan chain or adjacent glucan chain (Tetlow and Emes, 2014). BE (Branching enzymes) and DBE (Debranching enzymes) are antagonist in action wherein BE are involved in the introduction of α 1, 6 branching points whereas DBE trims improper branches generated by BE. Under the CAZy classification system, starch branching enzymes fall into two families *viz.,* Glycoside Hydrolase family, GH13 and GH57 (Lombard *et al.*, 2014). The two classes differ in the amino acid sequences and the tertiary structure. GH13 family is again subdivided into two subfamilies GH13_8 and GH13_9. GH13_8 comprises of prokaryotic enzymes whereas GH13_9 comprises of eukaryotic enzymes (Blesak and Janecek, 2012; Sawada *et al.*, 2014; Suzuki and Suzuki, 2016). In plants, branching enzymes are classified under the GH13 family. GH13 is grouped into two classes namely the type1/B family and type2/A family (Burton *et al.*, 1995). Type 1 class of branching enzymes utilizes amylose as the substrate and is involved in the transfer of long glucan chains whereas type 2 class of branching enzymes are involved in the transfer of short glucan chains making use of amylopectin as substrate. Nakamura *et al.* (2010) conducted *in vitro* analysis and structural characterisation of the products obtained when the recombinant branching enzymes was mixed with either amylose or amylopectin molecules as substrates.

In *Arabidopsis thaliana*, only two types of branching enzymes are involved in transitory starch degradation namely BE2.1 and BE2.2. BE2.1 and BE2.2 are the products of the genes *AtBE3* and *AtBE2* respectively (Fisher *et al.*, 1996). Dumez *et al.* (2006) reported that the absence of either of the BE2.1 or BE2.2 resulted in a slight change in the amylopectin structure and there is a collapse in the starch biosynthesis process in double defective mutants.

Cereal crops have groups of starch branching enzymes namely, *SBEI, SBEIIa* and *SBEIIb*. Biochemical studies revealed that SBEI exclusively branches longer chains whereas *SBEII* isoforms are involved in the transfer of shorter chains (Tetlow and Emes, 2014). Studies on the deletion of starch branching enzyme I (*SBE I*) revealed that it didn't contributed significantly in the reduction of starch content as well as didn't affected the kernel morphology (Regina *et al.*, 2015). Xia *et al.* (2011) investigated on the structure of starch kernels in rice and maize and detected minor changes in the fine structure of amylopectin molecule. *SBEIIb* mutations in maize and rice are commonly called as amylose-extender (ae) mutants. Regina *et al.* (2010) reviewed on the inhibition of both *SBEIIb* and *SBEIIa* in barley and wheat and concluded that *SBEIIa* inhibition increased the amylopectin content whereas *SBEIIb* inhibition resulted in minor change in the amylopectin content.

The transcriptome expression patterns of starch branching enzymes affect the synthesis of amylopectin, and consequently influence the morphology of starch granules. Carciofi *et al.* (2012) reported that the inhibition of *SBE I, SBEIIa* and *SBEIIb* in barley grains resulted in increased levels of amylose, generating high amylose starch that was readily resistant to enzymatic digestion. An increase in the amylose content from 25% to 75% and was observed in wheat due to the silencing of *SBEIIa* and *SBEIIb*. The enhanced expression of branching enzyme lead to a substantial increase in the branch numbers of amylopectin in the transgenic lines of rice (Kim *et al.*, 2005).

2.4.4 Starch Debranching enzyme – Pullulanase (3.2.1.41) and Isoamylase (3.2.1.68)

Debranching enzymes (isoamylase or pullulanase) are class of enzymes that selectively hydrolyze alpha 1, 6 glycosidic linkages producing low molecular weight linear short chain molecules. The starch debranching enzymes include both the pullulanase and isoamylase. Generally, pullulanase is involved in the hydrolysis of alpha 1, 6 glycosidic linkages in pullulan, limit dextrins and branched oligosaccharide yielding maltose, maltotriose and linear oligosaccharide respectively (Doman-pytka and Bardowski, 2004). The starch debranching enzymes from pullulanase contain a large number of linear oligosaccharides and higher yield when compared with isoamylase.

The short linear chains produced from the Debranching enzyme were used in the synthesis of spherical nanoparticles with dimension ranging between 30-45 nm through the self-assembly of the short chains (Qiu *et al.*, 2016). These nanoparticles were employed in the manufacture of economical and highly biocompatible delivery systems (Arijaje *et al.*, 2015). The sphere morphology of nanoparticles is achieved through spray drying (Zeng *et al.*, 2016).

Higher plants have four types of debranching enzymes. These include three types of isoamylase namely *ISA1*, *ISA2*, and *ISA3* and one pullulanase enzyme, *PUL* (Hussain *et al.*, 2003). *ISA1*is involved in the formation of two types of polymeric complexes. These are *ISA1*-homomer and *ISA1-ISA2* heteromer. *ISA1-ISA2* hetromer is present in a wide variety of plant species such as potato tuber (Lin *et al.*, 2012), *Arabidopsis thaliana* (Wattebled *et al.*, 2005), rice endosperm (Utsumi *et al.*, 2011) and maize endosperm (Kubo *et al.*, 2010). *ISA1*-homomer has a role in the starch biosynthesis process and is in rice endosperm (Fujita *et al.*, 2009), maize endosperm (Kubo *et al.*, 2010) and maize leaf (Lin *et al.*, 2013).

Morris and Morris (1939) studied the impact of defective debranching enzymes in plants and found that the starch granules were partly accumulated with

water-soluble glucans. A water-soluble glucan named phytoglycogen have been reported which is devoid of a proper cluster structure (Sumner and Sommers, 1944). The phytoglycogen-producing mutation is caused by a defect in the Isoamylase type I gene in various plants such as maize (Dinges *et al.*, 2003), rice (Utsumi *et al.*, 2011), potato (Bustos *et al.*, 2004) and *Arabidopsis thaliana* (Delatte *et al.*, 2005). Lin *et al.* (2012) reported the double mutants in maize which was defective in *ISA2* and *SSIII* produced WSG (water-soluble glucans) in the endosperm, although single mutants of either *ISA2* or *SSIII* could synthesize normal amylopectin.

In cereals, sugary-1 mutants of maize were retrieved as recessive homozygous sugary lines and have been marketing under the name "sweet corn" (James *et al.*, 2015). Isoamylase I (*ISAI*) was the sugary-1 gene that was present in maize (James *et al.*, 2015) which was very important for amylopectin biosynthesis in the case of maize endosperm.

2.4.5 Starch phosphorylase (2.4.1.1)

Starch phosphorylase or glycogen phosphorylase represents one of the largest families of enzymes involved in carbohydrate metabolism. The basic mechanism behind which starch phosphorylase works is by the transfer of sugar moieties from the donor to the specific acceptor molecules (Breton *et al.*, 2006). These enzymes are a part of Glycosyltransferase family 35 (GT35). Starch phosphorylase is widely distributed among different kingdoms comprising of plants, animals and prokaryotes (Nighojkar and Kumar, 1997). Two diverse forms of starch phosphorylase have been reported so far in all the plant varieties namely cytosolic and plastidial. Starch phosphorylase enzymes also take part in the synthesis of primers that are essential for the reactions catalyzed by starch synthases (Satoh *et al.*, 2008).

2.5 STARCH BIOSYNTHESIS PATHWAY RECONSTRUCTION

2.5.1 Comparative genomics approach

Comparative analysis of a large and numerous diverse sequenced genomes is revolutionizing the pace of gene discovery. The modern techniques of comparative genomics can drastically reduce the efforts that would be required to carry out manual sequencing and other tedious processes. Combining multiple new techniques in comparative genomics is often referred to as genome context analysis. A common theme of these efforts is the incorporation of various types of genomic evidences, like clustering of genes on the chromosome (Overbeek *et al.*, 1999), protein fusion events (Enright *et al.*, 1999; Marcotte *et al.*, 1999), signature profiles (Pellegrini *et al.*, 1999) and the presence of co-regulatory sites (Manson and Church 2000) to infer functional coupling for proteins participating in similar biological and cellular processes. Employing these techniques in the analysis of genes distributed on a genome produces valuable inferences (Makarova *et al.*, 2002).

2.5.2 Pathway Reconstruction

The plant-specific metabolic pathways have continuously been reconstructed not only for the wide purposes as found in the public repositories but also for a particular pathway investigation as demonstrated in various publications (Saithong *et al.*, 2013). Besides the model plant, Arabidopsis (Poolman *et al.*, 2009), the metabolic pathway reconstruction has been done in various commercially important crops such as wheat (Dupont, 2008), potato (Poolman *et al.*, 2009) and maize (Friso *et al.*, 2010). Only broad metabolic pathways have been reported so far in cassava (Sakurai *et al.*, 2007). Various techniques had been in use for investigating the starch biosynthesis process in cassava (Ihemere *et al.*, 2006; Friso *et al.*, 2010).

Sakurai *et al.* (2007) described the first metabolic pathways of cassava the wherein comparative genomic analysis of the full-length cDNA library had been exploited to the fullest even before the draft of cassava genome was released.

Two species with a well-defined genome organization and share an evolutionary resemblance with cassava such as castor bean and poplar (Djerbi *et al*., 2005) were selected for the study. Together with a whole set of genes from *Arabidopsis thaliana*, the cassava genome was studied with the help of comparative genomics.

In 2009, with the release of the cassava genome sequence (Goodstein *et al*., 2011), cassava research witnessed a dramatic revolutionary change. Rongsirikul *et al*. (2010) developed a reconstruction protocol of starch biosynthesis pathways in cassava making use of two template plants for the study. The template plants used were arabidopsis and potato. Nucleotide BLAST (BLASTn) between the templates and cassava were performed due to the scarcity in the data availability of cassava. The nucleotide sequences for both Arabidopsis and potato were retrieved from KEGG (Kanehisa and Goto, 2000). The sources of nucleotides being used for the study were genomic sequence, full length cDNA ESTs and partially sequenced ESTs. The cassava genome covers almost 95% with known genes. Secondly, the full length cDNA sequences were obtained from cDNA ESTs from leaves and roots of the tubers in response to extreme conditions like heat, stress, drought, post-harvest physiological deterioration which was reconstructed by Sakurai *et al*. (2007). Finally, partial sequences were derived from the partial ESTs from cassava. The use of inverse genomics approach also shed light in the work of Rongsirikul *et al*. (2010) where in the cassava genome was used as a database for search instead of being a query. The resulting pathway consisted of 35 biochemical reactions, 37 enzymes and 33 metabolites. Also the similarity search helped to identify at least 34 enzymes in cassava starch synthesis pathway. The reconstructed pathway of starch synthesis by Rongsirikul *et al*. (2010) covered two metabolic gaps that were present in the previous reconstruction protocol (Sakurai *et al*., 2007). The use of two template plants increased the comprehension of starch synthesis pathway in cassava in the aspect of cassava being similar to potato and not Arabidopsis. In a total of 34 enzymes present, 9 were identified only with potato indicating the evolutionary closeness of cassava and potato. However, only 4 enzymes were identified when a single template plant (Arabidopsis) was used.

Saithong *et al.* (2013) was able to the fill the pathway holes and overcome the shortages the previous pathway reconstruction protocol possessed (Rongsirikul *et al.*, 2010). The genetic variability of cassava was explored and that information was utilized to breed for a novel cassava line before the cassava genome data was released (El-Sharkawy, 2006). Nowadays, various high throughput technologies have been employed to disentangle the complex regulation underlying the biological processes in cassava plants, including cassava starch metabolism (Sakurai *et al.*, 2007; Li *et al.*, 2010; Yang *et al.*, 2011). The availability of sophisticated technologies to identify the genes and its expression like microarray, RNA-seq and NGS had contributed in studying the complex regulation controlling the behaviour of cells. The starch biosynthesis pathway was reconstructed with the aid of multiple template plants like arabidopsis, maize, rice, castor bean and potato. To get a better understanding on the starch and sucrose biosynthesis in cassava, the transcriptome data (Yang *et al.*, 2011) was integrated into the reconstructed pathway. Also manual curation and literature based validation added quality to the reconstructed frame. The proposed reconstruction pathway made use of parsimonious reconstruction framework and each step from collection till validation improved the comprehension of the reconstructed pathway than that of Sakurai *et al.* (2007) and Rongsirikul *et al.* (2010).

2.5.3 Microarray data integration

Yang *et al.* (2011) studied on the differences in gene expression of developing storage root (DR), fibrous root (FR) and mature storage root (MR) in order to gain a better understanding of cassava root development at the transcriptional level. The gene expression analyses were done for a total of 25 pathways and it was found that the gene expression was more in glycolysis/gluconeogenesis. The gene expression data of the rate limiting enzymes in the glycolysis and sucrose and starch metabolism were collected and mapped into the pathway to demonstrate the transcriptional regulation of genes

participating in the cassava root development. The enzymes serve as potential candidate targets in improving the starch yield (Yang *et al.*, 2011).

The reconstruction protocol of starch synthesis by Saithong *et al.* (2013) possesses more merits than that of Yang *et al.* (2011) in two aspects. Firstly, the reconstruction was carried out with the aid of five template crops viz., Arabidopsis, rice, maize, castor bean and potato. The KEGG metabolic pathway database contains reference pathways for starch and sucrose metabolism than a cassava specific pathway. The reconstructed pathway covers the reaction of cell wall synthesis (EC 3.2.1.27) which was not present in the work conducted by Yang *et al.* (2011). Secondly, the core focus of research was carbon partitioning in carbon metabolism cycle in the case of cassava root development so that a detailed inference could be drawn on these processes at the developmental stage of roots. AGPase (EC 2.7.7.27) activity in the starch biosynthesis pathway has enhanced to a larger extent during cassava root development which show similar results as of cereal endosperm (Emes *et al.*, 2003) and potato tuber (Hawker *et al.*, 1979).

MATERIALS AND METHODS

3. MATERIALS AND METHODS

The study entitled "Comparative and Functional Genomics analysis of Starch biosynthesis pathways in Cassava" was carried out at the Section of Extension and Social Sciences, ICAR-Central Tuber Crops Research Institute, Sreekariyam, Thiruvananthapuram during 2016-2017. In this chapter, details regarding experimental materials and methodology used in the study are elaborated.

3.1 BLAST TEMPLATE PREPARATION

The nucleotide and protein sequences involved in carbon-dioxide fixation cycle (Calvin cycle), starch and sucrose synthesis in cassava and other template plants like arabidopsis, maize, rice, potato and castor bean were retrieved from their respective genome databases. The databases used in the study were Phytozome (Goodstein *et al.*, 2011), TAIR (Rhee *et al.*, 2003), Gramene (Monaco *et al.*, 2014), Rice Genome Annotation Project (Yuang *et al.*, 2003), Sol Genomics Network (Fernandez-Pozo *et al.*, 2015) and KEGG (Kanehisa and Goto, 2000) for cassava, Arabidopsis, maize, rice, potato and castor bean respectively. The template plants were selected in having a distinct genome organization, being a starch crop and also having evolutionary and physiological resemblance with cassava.

3.1.1 Reciprocal BLASTp

Reciprocal BLASTp (Altschul *et al.*, 1990) is used for the *in silico* prediction of the putative orthologues in an unknown organism by means of bidirectional sequence alignments. Reciprocal BLASTp progresses through two steps. In the first step, a particular protein sequence derived from a template plant was used as query to compare against the cassava protein library which is called as the first BLASTp. Consequently, the cassava proteins from the BLASTp were used to align against the template plant protein library. Based on a lower E-value

and a higher query coverage (>= 80 %), those cassava sequences that met the given set critical value were taken for further analysis.

3.1.2 Function assignment

E2P2 (Ensemble Enzyme Prediction Pipeline) Version3.0 (Chae *et al.*, 2014) which is a subsidiary tool of PMN (Plant Metabolic Network) was used for functional annotation of the electronically predicted sequences. The Ensemble Enzyme Prediction Pipeline version 3.0 is composed of Enzyme Function (EF) classes containing four-part Enzyme Commission numbers and MetaCyc reaction identifiers. Using the E2P2 software, species- specific metabolic databases can be created. E2P2 methodologically assemble results from two molecular function annotation algorithms using a prescribed scheme of Ensemble. For a whole genome of an organism, all the protein sequences are manually provided as individual subsets against the base-level annotation scheme. Finally, with the aid of maximum weighted integration algorithm, the base-level predictions are compiled to final annotation network. The annotation of sequences by E2P2 is by homology transfer through integration of single and multiple sequence models having enzymatic function. E2P2 is a versatile tool which is able to run on a 64 bit Linux platform which is packed with various dependencies for performing diverse functions. The input file is loaded in the FASTA format and it should be a protein sequence. The work flow is given in Fig. 1.

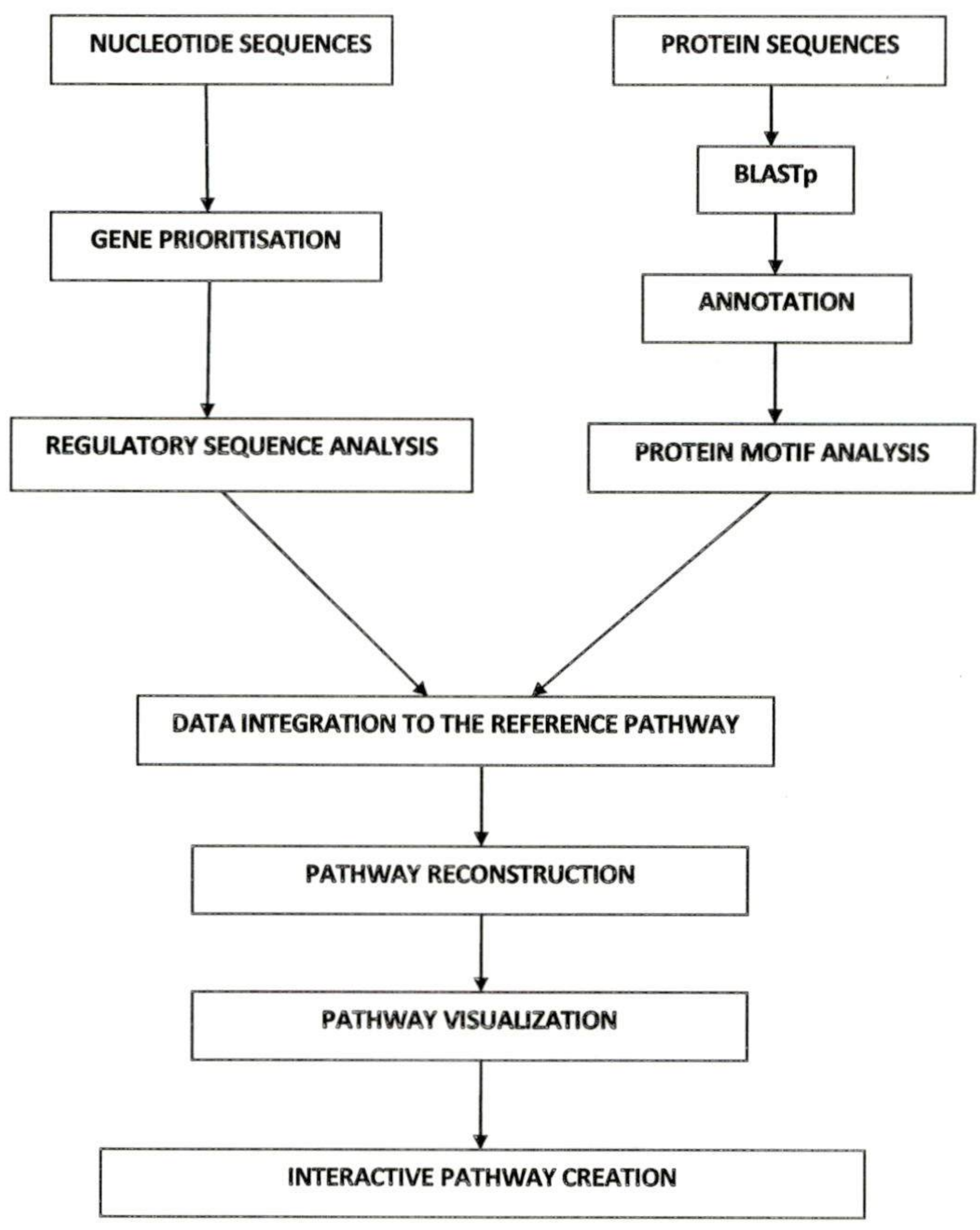

Fig. 1 Workflow for the starch biosynthesis pathway reconstruction in cassava

3.2 PROTEIN MOTIF MINING AND ANALYSIS

Motif mining in the protein and nucleotide sequences is one among the decisive step in protein and nucleotide sequence analysis. Motifs in the context of proteins are those related sequences or sub-sequences that are preserved all along the course of evolution of protein. For a DNA, motif could be a transcription factor binding site. Motifs do occur multiple times in a sequence.

3.2.1The MEME suite version 4.12.0

The predicted gene sequences were subjected to motif analysis. The Meme Suite (Timothy *et al.*, 2009) contains various tools for motif discovery, motif enrichment, motif scanning and motif comparison.

3.2.1.1 Motif discovery – MEME

The tool used for motif discovery was MEME (Multiple Em for Motif Elicitation) (Timothy *et al.*, 2009). When a sequence or a set of sequences are provided, MEME uses *De Novo* approach for the discovery of novel ungapped motifs. The motifs were represented as Position-dependent-letter-probability matrices based on statistical modelling.

3.2.1.2 Motif enrichment – GOMO

The tool that was used for motif enrichment was GOMO (Gene Ontology for Motifs) (Fabian *et al.*, 2010). GOMO looks for protein motifs that are associated with some genes linked to one or more genome ontology terms. The biological functions of the motifs could be elucidated from significant GO (Genome Ontology) terms.

3.2.1.3 Motif scanning – MAST

The tool used for motif alignment was MAST (Motif Alignment and Search Tool) (Timothy *et al.*, 2009). MAST looks for sequences that matches to a set of motifs provided as the query by the user and sort sequences based on the best combined match to all motifs.

3.2.1.4 Motif comparison – TOMTOM

Motif comparison was carried out using TOMTOM (Shobit *et al.*, 2007). It compares one or more motifs against a database containing a set of known motifs. TOMTOM rank the motifs that are present in the database and create an alignment for each significant match produced.

3.2.2 Protein-protein interaction studies

STITCH version 5.0 (Kuhn *et al.*, 2014) was used to study about the protein-protein interactions. It is a data integration tool wherein the data which is dispersed in the literature and various other databases of biological pathways are brought in a nutshell and henceforth an interaction network is created with a high confidence value. The network shows two types of interconnecting lines *viz.*, thicker and thinner lines. Thicker lines ensure comparatively a higher degree of interaction when compared with the thinner fade lines.

3.3 REGULATORY SEQUENCE ANALYSIS

The identification and also the analysis of regulatory elements that are crucial for gene expression could be greatly facilitated with the aid of database-assisted bioinformatics approaches. Database assisted Promoter analysis is one of the ways to study gene regulatory networks. With the advancements in sequencing technology and also the availability of whole genome sequences for most of the plant species, a number of promoter databases have been developed to study transcriptional regulation.

3.3.1 Promoter analysis

The promoter analysis was performed using the PROSCAN version 1.7 (Prestridge, 1995). PROSCAN is maintained by the Bioinformatics and Molecular Analysis Section (BIMAS). PROSCAN reads through the primary input sequence and look for putative eukaryotic Pol II promoter sequences in the query.

3.3.2Transcription factor binding sites (TFBS) prediction

The identification of putative transcription factor binding sites was carried out using ALGGEN-PROMO, JASPAR and Plant Transcription Factor Database (PlantTFDB). The identification of transcription factor binding sites help in understanding the strength and regulation of the promoters which in turn enhance gene expression. Most of the transcription factors possess a DNA- binding domain which is very much essential in the recognition of short elements or sequences present in the promoter. Transcription factors play a vital role in gene regulation and there are several databases that are exclusively dedicated to them. Plant Transcriptional Regulatory Network is required in understanding the distinct functional and evolutionary features underlying novel transcription factors.

3.3.2.1 ALGGEN-PROMO

PROMO (Xavier *et al.*, 2002) is a virtual laboratory used for the prediction of transcription factor binding sites in the gene sequences in particular the DNA sequences from a species or a group of desired species. The TFBS prediction is based on the construction of specific binding site matrices. The output can be downloaded in both textual as well as graphical interface.

3.3.2.2 JASPAR

JASPAR (Anthony *et al.*, 2013) is an open access database consisting of experimentally defined transcription factor binding profiles in eukaryotes. Compared with other resources that serve the same purpose, JASPAR has an add-on advantage in providing high quality and non-redundant output. Another feature of JASPAR is the dynamic clustering of motifs which is done using STAMP tool.

3.3.2.3 Plant Transcription Factor Database (PlantTFDB)

Plant Transcription Factor Database (Jin *et al.*, 2014) is a comprehensive platform of transcription factors, regulatory elements and their functional enrichment analyses which serve as a better tool in the creation of advanced plant transcriptional regulatory networks. TF prediction, Binding site prediction,

Regulation prediction, Gene Ontology (GO) Enrichment, TF Enrichment are some of the tools in the database.

3.4 GENE PRIORITISATION

The tool used for candidate gene prioritisation was ToppGene Suite (Chen *et al.*, 2009). ToppGene Suite works based upon functional annotation and protein interaction network. ToppGene Suite is a package composed of various tools like ToppFun for functional enrichment of gene, ToppGene for candidate gene prioritisation, ToppNet for ranking the prioritised genes based on topology and ToppGenet for prioritising the maximum likelihood genes in a Protein-protein interaction network. RiceNet (Lee *et al.*, 2015) is a rice probabilistic functional network which was created by linking genes that are functional altogether in *Oryza sativa* and used a modified Bayesian algorithm in the construction of the network.

3.5 CONSOLIDATED PATHWAY CONSTRUCTION

Carbon-dioxide fixation pathway or the Calvin-Benson-Bassham cycle and sucrose biosynthesis II were linked with the starch biosynthesis pathway in cassava to create a common consolidated pathway using a unified intermediate common to all the pathways. The pathway of starch biosynthesis was recreated using Cell Illustrator version 5.0 (Nagasaki *et al.*, 2010). It is the software employed by biologists to draw, model and stimulate complex biological processes. This can also be used for modelling metabolic pathways, gene regulatory pathways and signal transduction reactions.

3.6 DATA INTEGRATION

The outputs from various analyses were incorporated into the pathway constructed to gain a better understanding of starch biosynthesis in cassava. An insight into the molecular mechanism underlying a particular enzyme was obtained through various analyses *viz.,* candidate gene prioritisation, regulatory elements analysis and protein motif analysis. Protein-protein interaction networks were also created.

3.6.1 Pathway reconstruction and Visualization

A descriptive platform was constructed using Cell Illustrator version 5.0 (Nagasaki *et al.*, 2010). An interactive platform for the reconstruction and also the visualization of starch biosynthesis pathways in cassava was created using VANTED version 2.6.3 (Hendrik *et al.*, 2012). VANTED (Visualization and Analysis of Networks containing Experimental Data) is a Java based extendable network visualisation and analysis tool with focus on applications on life sciences. It allows users to create and edit networks as well as mapping experimental data onto networks. VANTED has got inbuilt statistical algorithms for the fast evaluation and interpretation of the mapped data. Experimental datasets are visualized on VANTED as graphical charts which display time series, different treatment applications as well as environmental conditions underlying a biological process.

3.6.2 Microarray Data Integration

A microarray measures the gene expression of thousands of genes simultaneously. It has revolutionized the field of biological research in discovering a huge set of genes whose expression levels determine a particular cell type, disease, treatment or a developmental stage. Expression data is now available for more than 100 species (Chen *et al.*, 2008). The public data repositories for the collection, annotation, dissemination and redistribution of the microarray gene expression data include NCBI GEO (Edgar *et al.*, 2002), EBI ArrayExpress (Rocca-Serra *et al.*, 2003) and NIG CIBEX (Ikeo *et al.*, 2003).

The differentially expressed genes from different types of roots in cassava namely the fibrous roots, matures roots and developing storage roots were selected for the integration of transcriptome data into the reconstructed biochemical pathway of starch synthesis. The gene expression data for cassava roots (Yang *et al.*, 2011) was retrieved from the NCBI GEO (database and mapped into the reconstructed pathway of starch biosynthesis. The dataset used for the study was GSM25813 series matrix file. 20,840 genes were present in the

microarray and after normalisation using GeneXplain platform, data was reduced to 5000 genes. Using MeV (Multi experiment Viewer), heat map was generated and clustering was performed. The SNR (Sound to Noise ratio) was set as 2.58 and P-value was kept at less than 0.05 to get significant genes present in the microarray.

Differentially expressed genes were identified from the heat map and after statistical analysis like HCL clustering (Hierarchical clustering), SAM (Significance Analysis of Microarray) graph and Volcano plot. The differentially expressed genes were then mapped into the reconstructed pathway and viewed as an interactive pathway on VANTED.

RESULTS

4. RESULTS

The results of the study "Comparative and Functional Genomics Analysis of Starch Biosynthesis Pathways in Cassava" carried out at the Section of Extension and Social Sciences, ICAR-Central Tuber Crops Research Institute, Sreekariyam, Thiruvananthapuram during 2016-2017 are presented in this chapter.

4.1 BLAST TEMPLATE PREPARATION

The template plants used for the study were arabidopsis, maize, rice, castor bean and potato. The starch related genes in the cassava were identified from their respective orthologues present in the template plants. Starch related genes along with the protein sequences of all the template plants were retrieved from their respective databases. TAIR (https://www.arabidopsis.org/), Gramene (http://ensembl.gramene.org/Zea_mays/Info/Index/), Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/analyses_search_blast.shtml), KEGG (http://www.genome.jp/kegg/pathway.html) and Sol Genomics Network (https://solgenomics.net/) were used for the retrieval of protein sequences for Arabidopsis, maize, rice, castor bean and potato respectively. The entire genome information for cassava was taken from the Phytozome database (https://phytozome.jgi.doe.gov/pz/portal.html). Additionally, PMN (Plant Metabolic Network) and KEGG (Kyoto Encyclopaedia of Genes and Genomes) were used in the further dissemination of information involved in the starch synthesis. The starch and sucrose metabolism pathway (map00500) and Carbon-dioxide fixation/ C_3 cycle in photosynthetic organisms (map00710) were taken from the KEGG database to serve as the reference pathways for the experiment. The genes involved in carbon dioxide fixation/Calvin cycle (Table 1), sucrose synthesis I (Table 2) sucrose synthesis II (Table 3) and starch synthesis (Table 4) along with its sub-cellular localisation was retrieved from KEGG.

Table 1. Genes involved in carbon-dioxide fixation in cassava

Gene	Localisation	EC Number
Ribulose bisphosphate carboxylase small chain	Chloroplast	4.1.1.39
Phosphoglycerate kinase	Extracellular	2.7.2.3
Glyceraldehyde3 phosphate dehydrogenase	Non-cytoplasm	1.2.1.12
Triose-phosphate isomerase/ mutase	Cytoplasm	5.3.1.1
Sedoheptulose-bisphosphatase	Chloroplast	3.1.3.37
Transketolase	Peroxisome	2.2.1.1
Ribose-5-phosphate isomerase	Cytoplasm	5.3.1.6
Fructose-bisphosphate aldolase	Cytoplasm	4.1.2.13
Fructose 1,6 bisphosphatase	Chloroplast	3.1.3.11
Ribulose-phosphate -3 epimerase	Chloroplast	5.1.3.1
Phosphoribulokinase	Chloroplast	2.7.1.19

Table 2. Genes involved in sucrose synthesis I in cassava

Gene	Localisation	EC Number
Phosphoglycerate kinase	Extracellular	2.7.2.3
Glyceraldehyde 3 phosphate dehydrogenase	Cytoplasm	1.2.1.12
Fructose bisphosphate aldolase	Cytoplasm	4.1.2.13
Fructose 1,6 bisphosphatase	Chloroplast	3.1.3.11
Glucose-6-phosphate isomerase	Cytoplasm	5.3.1.9
Phosphoglucomutase	Extracellular	5.4.2.2
UDP glucose pyrophosphorylase 3	Extracellular	2.7.7.9
Glycosyltransferase / sucrose phosphate synthase	Nucleus	2.4.1.14
Sucrose-6 phosphate phosphatase	Extracellular	3.1.3.24

Table 3. Genes involved in sucrose synthesis II in cassava

Gene	Localisation	EC Number
Alpha ,4 glucanotransferase	cytoplasm	2.4.1.25
Phosphoglucomutase	Extracellular	5.4.2.2
PGI(glucose-6-phosphate isomerase)	cytoplasm	5.3.1.9
UDP glucose pyrophosphorylase	Extracellular	2.7.7.9
Glycosyltransferase	Nucleus	2.4.1.14
Sucrose-6-phosphate phosphatase	Extracellular	3.1.3.24
Hexokinase	Chloroplast	2.7.1.1
Glycogen phosphorylase	Extracellular	2.4.1.1

Table 4. Genes involved in starch synthesis in cassava

Gene	Localisation	EC Number
PGI (glucose-6 phosphate isomerase)	Cytoplasm	2.4.1.25
Phosphoglucomutase	Extracellular	5.4.2.2
Glucose-1-phosphate adenyl transferase large subunit 3	Cytoplasm	5.3.1.9
Starch synthase 2	Extracellular	2.7.7.9
Alpha 1,4 glucan branching enzyme	Nucleus	2.4.1.14
Isoamylase 1	Extracellular	3.1.3.24
Isoamylase 2/ starch branching enzyme 2	Chloroplast	2.7.1.1
Isoamylase 3	Extracellular	2.4.1.1
NDP-glucose-starch glucosyltransferase/ waxy protein	Chloroplast	2.4.1.242

Reciprocal BLASTp was performed in order to obtain the putative reciprocal BLAST hits (RBHs). The strategy followed is inverse comparative genomics approach where in the cassava genome is used as a search database instead of being a query. After Reciprocal BLASTp, three putative protein sequences were obtained with a lower E-value and a very high query coverage ranging from 95-99%. The obtained sequences were then passed through E2P2 (Ensemble Enzyme Prediction Pipeline) software for function assignment.

E2P2 predicts and annotates functions to the electronically predicted sequences without any human oversight. E2P2 version 3.0 was used for assigning function to the identified sequences. The Ensemble Enzyme Prediction Pipeline version 3.0 is composed of Enzyme Function (EF) classes containing Enzyme Commission numbers and MetaCyc reaction identifiers as shown in Table 5. Using the E2P2 software, species- specific metabolic databases could be created. The annotation of sequences by E2P2 is done by homology transfer through integration of single and multiple sequence models having enzymatic function.

Table 5. Function assignment to the predicted sequences using E2P2 tool

Gene ID	BLAST	PRIAM	Annotated function	MetaCyc ID
Manes.03G198900	EF00058	EF00058	Sucrose phosphate synthase	RXN 14058
Manes.09G059400	EF11550	EF11550	Alpha 1,4 branching enzyme	RXN 14372
Manes.03G044400	EF11705	EF05788	UDP glycoprotein superfamily 13 (UDP-Glycosyltransferase)	RXN 14378

4.2 PROTEIN MOTIF ANALYSIS – The MEME suite

The sequences predicted were subjected to protein motif analysis in the MEME suite version 4.12.0. The MEME suite consists of various tools for motif discovery, motif scanning, motif enrichment and motif comparison.

4.2.1 Motif discovery – MEME

MEME uses *De Novo* approaches in the prediction of ungapped motifs. The motifs were represented as Position-dependent-letter-probability matrices based on statistical modelling. Three motifs were discovered in genes with predicted function as that of alpha 1, 4 glucan branching enzyme from MEME tool as shown in Fig 2. Similar prediction was carried out using the other two predicted sequences. These motifs were then subjected to various analyses like motif enrichment, motif scanning and motif comparison.

4.2.2 Motif enrichment – GOMO

GOMO searches for protein motifs which are in collaboration with some genes that are linked to one or more genome ontology terms. The biological functions of the motifs could be elucidated from significant GO (Genome Ontology) terms. The output of MEME served as the input for GOMO. The motifs obtained when putative sequences were loaded into MEME were then provided to GOMO for functional annotation. Nine predictions were obtained for putative sequence with annotated function of alpha 1, 4 glucan branching enzyme as shown in Table 6. From the table, it was inferred that majority of protein motifs can act as transcription factors and also assist in various biosynthetic processes in plants like Xanthophyll biosynthesis. Being the accessory pigments in the photosynthesis process, xanthophylls play a vital role in plants and hence the protein motifs linked to xanthophyll biosynthesis are crucial in plant metabolism.

4.2.3 Motif scanning – MAST

The tool used for motif alignment was MAST (Motif Alignment and Search Tool). MAST looks for sequences that matches to a set of motifs provided

as the query by the user and sort sequences based on the best combined match to all motifs. Motifs are represented as position-dependent probability (or scoring) matrices that describe the probability (or score) of each possible letter at each position in the pattern. Three motifs were given as input in MAST and after similarity check and alignment; three sequences were obtained which were similar to the input motifs obtained from MEME.

Table 6. Protein motifs identified linked to GO terms using GOMO tool

GO term	score	p-value	q-value	specificity	GO name
GO: 0009057	3.503e-06	2.652e-07	3.321e-04	20 %	chloroplast (CC)
GO: 0003677	4.828e-03	2.652e-07	3.321e-04	3 %	DNA binding (MF)
GO: 0003700	6.860e-03	2.652e-07	3.321e-04	83 %	transcription factor activity (MF)
GO: 0005739	8.755e-03	5.304e-07	4.981e-04	12 %	mitochondrion (CC)
GO: 0009941	1.220e-02	1.326e-06	9.962e-04	65 %	chloroplast envelope (CC)
GO: 0016123	1.661e-02	4.243e-06	2.657e-03	100 %	xanthophyll biosynthetic process (BP)
GO: 0005634	1.769e-02	5.304e-06	2.846e-03	2 %	nucleus (CC)
GO: 0005524	2.772e-02	4.004e-05	1.880e-02	100 %	ATP binding (MF)
GO: 0009535	2.967e-02	5.330e-05	2.225e-02	66 %	chloroplast thylakoid membrane (CC)

CC: Cellular Component, MF: Molecular Function, BP: Biological Process

4.2.4 Motif comparison – TOMTOM

TOMTOM compares one or more motifs against a database containing a set of known motifs. TOMTOM rank the motifs that are present in the database and create an alignment for each significant match produced. Eight motifs were obtained on providing the predicted sequence (α 1, 4 glucan branching enzyme) as input which is shown in the Table 7. From the table, it is clear that the motifs present in the query have transcriptional factor activity serving as the major transcription factor domains. The transcription factor domains present in the sequence include MYB, SRS, bZIP, bHLH *etc.*

Table 7. Ranking of motifs from alpha 1, 4 glucan branching enzyme using TOMTOM tool

Name	Database	p-value	E-value	q-value
MYB99	Arabidopsis DAPv1	4.91e-04	7.76e- 01	5.17e-01
MYB92	Arabidopsis DAPv1	2.58e-03	2.25e+00	5.17e-01
SRS7	Arabidopsis DAPv1	3.48e-03	3.04e+00	5.17e-01
bZIP50	Arabidopsis DAPv1	7.27e-03	6.34e+00	5.17e-01
Orphan AT1G24250	Arabidopsis DAPv1	8.75e-03	7.63e+00	5.17e-01
ANAC087	Arabidopsis DAPv1	1.12e-02	9.80e+00	5.17e-01
bHLH130	Arabidopsis DAPv1	9.71e-04	8.47e- 01	5.17e-01
MYB43	Arabidopsis DAPv1	6.03e-02	5.98e- 04	5.17e-01

4.3 PROTEIN-PROTEIN INTERACTION NETWORK

Protein-protein interaction network was created for predicted α1, 4 glucan branching enzyme and predicted UDP- glycoprotein super family using STITCH tool. The graphical outputs for α 1, 4 glucan branching enzyme and UDP-glycosyl- transferase are shown in Fig 3 and Fig 4. The network is created by cross-comparison of shared interactions from different literatures and also by natural language processing. In the figure, red coloured nodes represent query proteins which are the first shell of interactors where as white nodes indicate second shell of interactors. Similarly, small nodes graphically resembles proteins whose structures are not known or predicted where as large nodes indicate that the proteins have got some 3D structure which is either known to the user or predicted. The network created is a closed array consisting of various enzymes involved in starch biosynthesis like Isoamylase, starch branching enzyme, granule-bound starch synthase, soluble starch synthase, disproportionating enzyme *etc*.

4.4 REGULATORY SEQUENCE ANALYSIS FOR THE PREDICTED SEQUENCES

4.4.1 Promoter analysis

The promoter analysis was performed using the PROSCAN version 1.7 which is maintained by the Bioinformatics and Molecular Analysis Section (BIMAS). PROSCAN scans the query sequence and look for putative eukaryotic polymerase II sequences in the query. Based on scoring homology, PROSCAN creates a textual output. The output of predicted α 1, 4 glucan branching enzyme is shown in Table 8. The table shows five transcriptional signals predicted to the upstream of the query sequence and seven transcriptional signals to the downstream. Most of the predicted signals have transcriptional regulatory roles and serve as DNA binding domains. Similarly, promoter analysis was carried out for the other two predicted sequences.

Table 8. Promoter region prediction in predicted alpha 1,4 glucan branching enzyme using PROSCAN tool

Name	Strand	Location	Weight
NF-KB	+	986	1.008000
ATF	-	969	1.157000
CREB	+	965	1.147000
E4F1	+	964	1.201000
ATF/CREB	+	964	1.138000
CREB	+	963	2.549000
TFIID	-	957	2.618000
TFIID	-	947	1.971000
TRF	-	843	2.151000
TFIID	-	754	1.971000
TFIID	-	754	2.920000
TFIID	-	754	2.618000

4.4.2 Transcription Factor Binding sites (TFBS) prediction

The identification of putative transcription factor binding sites was carried out using ALGGEN-PROMO, JASPAR and Plant Transcription Factor Database (PlantTFDB).

4.4.2.1 *ALGGEN-PROMO*

TFBS defined in the TRANSFAC database are used to construct specific binding site weight matrices for TFBS prediction. The visual output of predicted alpha 1,4 glucan branching enzyme is shown in Fig 5.

Discovered Motifs

	Logo	E-value	Sites	Width	More	Submit/Download
		9.0e+001	4	39		
		1.3e+001	4	43		
		4.9e+003	3	41		

Fig 2. Motifs identified using MEME tool

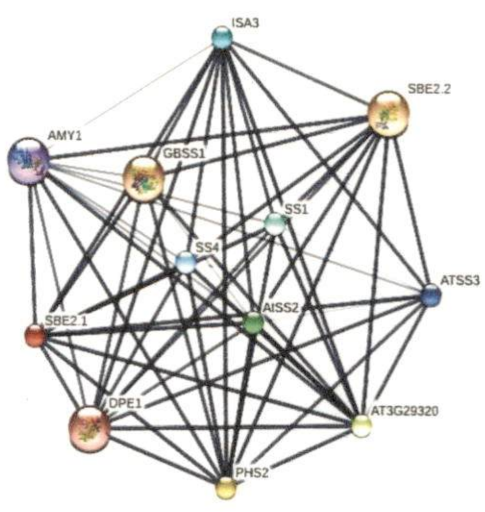

Fig 3. Protein-protein interaction network of predicted Starch synthase created using STITCH

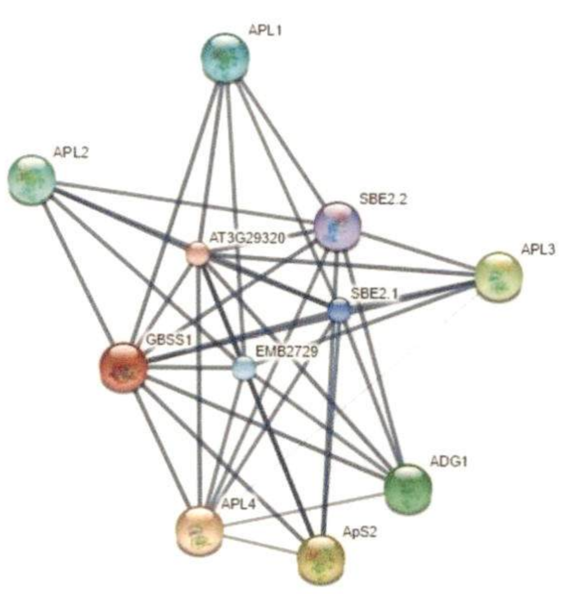

Fig 4. Protein-protein interaction network of predicted UDP-glycosyl transferase created using STITCH

4.4.2.2 *JASPAR*

JASPAR is an open access database consisting of experimentally defined transcription factor binding profiles in eukaryotes. JASPAR also creates a weight matrices based on the position of the putative transcription factors in the query sequence. JASPAR was used to determine the transcription factor binding sites in predicted sequences. The output of JASPAR for the predicted α 1, 4 glucan branching enzyme is shown in Table 9. Dof2 matrix was used as the transcription factor binding profile which was obtained from *Arabidopsis thaliana*. Being one of the template plant used for the study, Arabidopsis serve as a better option in the selection of transcription factor binding profile.

Table 9. Transcription factor binding sites predicted using JASPAR for putative α 1, 4 glucan branching enzyme

Model ID	Model Name	Score	Relative score	Start	End	Strand	Predicted sequence
MA0020.1	Dof2	5.518	0.840549163722296	105	110	1	AAAGGG
MA0020.1	Dof2	8.760	0.999993792767326	126	131	-1	AAAGCA
MA0020.1	Dof2	5.518	0.840549163722295	139	144	-1	AAAGGG
MA0020.1	Dof2	6.086	0.8684839414205	308	313	-1	AAAGTT
MA0020.1	Dof2	6.086	0.8684839414205	350	355	-1	AAAGGT
MA0020.1	Dof2	6.891	0.908074603475173	477	482	-1	AAAGAA
MA0020.1	Dof2	6.484	0.888057958187903	493	498	1	AAAGAT
MA0020.1	Dof2	6.484	0.888057958187903	501	506	-1	AAAGAT
MA0020.1	Dof2	8.354	0.980026328426709	509	514	-1	AAAGCT
MA0020.1	Dof2	5.518	0.840549163722296	536	541	1	AAAGTG

4.4.2.3 *Plant Transcription Factor Database (PlantTFDB)*

Plant Transcription Factor Database is a comprehensive platform of transcription factors, regulatory elements and their functional enrichment analyses which serve as a better tool in the creation of advanced plant transcriptional regulatory networks. Using, PlantTFDB, transcription factor binding sites were predicted in the sequences identified *in silico*. Nine predictions were obtained for α 1, 4 glucan branching enzyme on chromosome 17. The tabulated form of the output for the predicted α 1, 4 glucan branching enzyme is shown in Table 10. From the table, it was inferred that most of the transcription factors predicted belong to the AP2 family which has a substantial role in sucrose transport.

Table 10. Transcription factor binding site prediction using PlantTFDB for putative α 1, 4 glucan branching enzyme

Motif	Family	Sequence ID	Position	Strand	p-value	q-value	Matched sequence
Manes.13G001800	AP2	Chromosome 17	523-542	+	3.45e-05	0.078	GGAGACAAAAATGAAAGTGA
Manes.12G117800	AP2	Chromosome 17	386-406	+	6.31e-05	0.108	TGGCTCGAATAGCCAGTTGCA
Manes.13G001800	AP2	Chromosome 17	415-434	+	9.29e-05	0.106	CAGCCCAAGTGGAAGAGAAG
Manes.12G117800	AP2	Chromosome 17	864-884	_	9.84e-05	0.108	GGGCTCAAGCCTCCAGAGGCC
Manes.18G015400	ARF	Chromosome 17	667-676	+	8.50e-05	0.193	GCAGACAATA
Manes.17G047000	B3	Chromosome 17	323-330	_	2.46e-05	0.052	CTGCCTGC
Manes.13G043500	B3	Chromosome 17	322-330	_	6.92e-05	0.145	CTGCCTGCA
Manes.17G089900	BBR-BPC	Chromosome 17	429-452	+	7.91e-05	0.177	GAGAAGTGAAGCTGGGTTGAGAA
Manes.14G013300	bHLH	Chromosome 17	417-430	_	8.32e-06	0.019	TCTTCCACTTGGGC

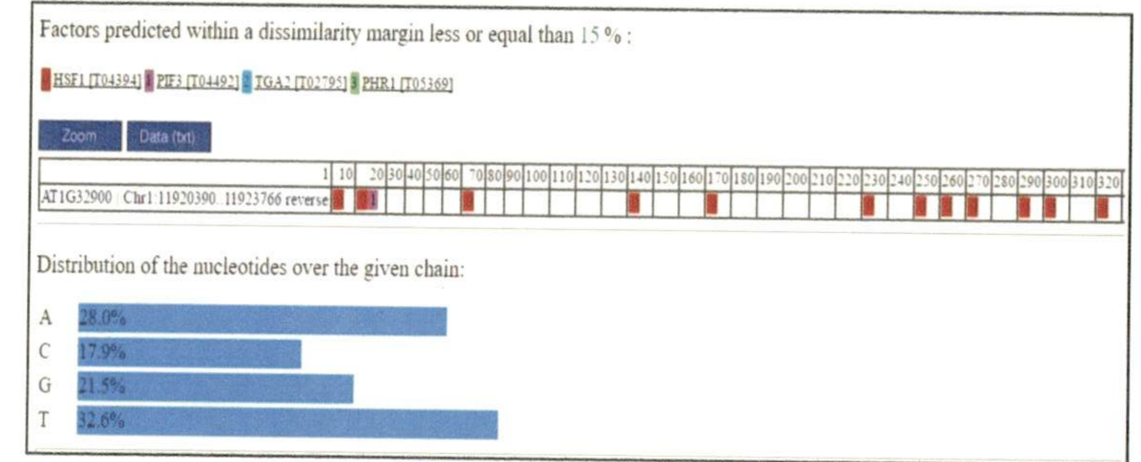

Fig 5. Transcription factor binding site prediction using Alggen-Promo in the predicted α 1, 4 glucan branching enzyme

4.5 GENE PRIORITISATION

The tool used for candidate gene prioritisation was ToppGene Suite. ToppGene Suite works based upon functional annotation and protein interaction network. ToppGene Suite is a package composed of various tools like ToppFun for functional enrichment of gene, ToppGene for candidate gene prioritisation, ToppNet for ranking the prioritised genes based on topology and ToppGenet for prioritising the maximum likelihood genes in a Protein-protein interaction network. The tool used for gene prioritisation in rice was RiceNet. Based on the literature review and comparative analysis in one of the template plant, Oryza, 5 genes were prioritised in cassava which can be used as candidate genes in starch biosynthesis pathway. One of the genes prioritised is a negative regulator in the starch biosynthesis pathway in cassava which has a role similar to plastidial adenylate kinase. It was reported that when the levels of plastidial ADK increases, the rate of starch biosynthesis was found to be retarded. Other genes identified assist in chromatin remodulation, root elongation, lncRNA metabolism etc. The candidate genes predicted using ToppGene suite is shown in Table 11.

4.6 CONSOLIDATED PATHWAY CONSTRUCTION

The pathway of starch biosynthesis was recreated using the Cell Illustrator version 5.0. The starch biosynthesis process in cassava composing of carbon-dioxide fixation, starch and sucrose biosynthesis were made into a common pathway along with the incorporation of various transcription factors identified and the predicted gene sequences. The recreated starch biosynthesis pathway in cassava is shown in Fig 6.

4.7 DATA INTEGRATION

The outputs from various analyses were incorporated into the pathway constructed to gain a better understanding of starch biosynthesis in cassava. These include *cis* regulatory elements like promoters and transcription factors predicted through *in silico* methods and also the candidate genes identified regulating starch biosynthesis in cassava.

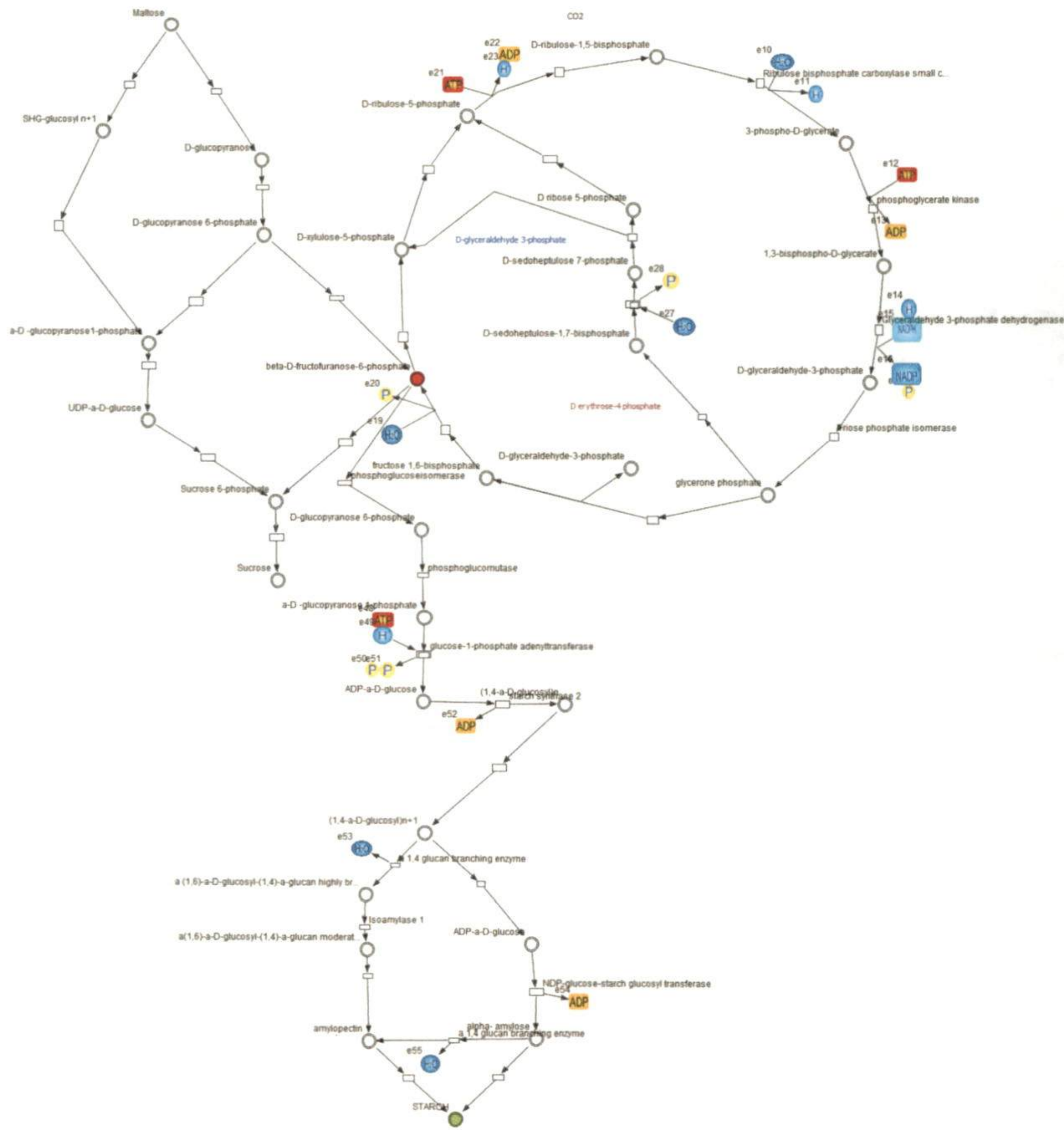

Fig 6. Consolidated pathway of starch biosynthesis in cassava constructed using Cell Illustrator

4.7.1 Pathway reconstruction and Visualization

An interactive platform for the reconstruction and also the visualization of starch biosynthesis pathways in cassava was created using the recent VANTED version 2.6.3. The reconstructed pathway which was viewed on VANTED platform is shown in Fig 7.

4.7.2 Microarray Data Integration

The differentially expressed genes of different types of cassava roots were retrieved out of the microarray data extracted from the NCBI GEO database and mapped into the reconstructed pathway of starch biosynthesis. The dataset used for the study was GSM25813 series matrix file. 20,840 genes were present in the microarray and after normalisation using GeneXplain platform, data was reduced to 5000 genes. The heat map of the genes after normalization is shown in Fig 8. Using MeV (Multi experiment Viewer), heat map was generated and clustering was performed. The SNR (Sound to Noise ratio) was set as 2.58 and P-value was kept at less than 0.05 to get significant genes present in the microarray. HCL clustering was done for the genes after normalization and identified the significant genes as shown in Fig 9. Twelve significant genes were identified which have a potential role in the starch biosynthesis pathway in cassava. The heat map of significant genes generated from MeV software is shown in Fig 10.

Differentially expressed genes were identified from the heat map and after statistical analysis like HCL clustering (Hierarchical clustering) SAM (Significance Analysis of Microarray) graph and Volcano plot. SAM graph is shown in Fig 11 and the volcano plot created for the significant genes is shown in Fig 12. Seven out of twelve significant genes were found to be differentially expressed in different growth stages. These are contig2017, TA9083_3983, TA5570_3983, contig5561, TA8522_3983, contig4441 and BM259732.

The transcriptome data integration showed that maximum gene expression was observed in the mature roots as compared to developing and storage roots. The important enzymes in the pathway of starch synthesis confirmed by microarray analysis include ADP glucose pyrophosphorylase (AGPase), starch branching enzyme and glucan phosphorylase. DR1 resembled FR whereas DR2 and DR3 resembled MR in terms of expression pattern.

The differentially expressed genes were tabulated as shown in Table 12. Finally, the differentially expressed genes were mapped into the reconstructed pathway and viewed as an interactive pathway on VANTED as shown in Fig 13.

Table 11. Candidate genes involved in starch biosynthesis in cassava predicted through ToppGene suite

Arabidopsis thaliana	Oryza sativa	Manihot esculenta	GO: Biological Process	EO: Enzyme Ontology
AT2G37250	LOC_Os03g03820	cassava4.1_013232	Negative regulator of starch biosynthesis, nucleotide phosphorylation, root development, shoot development	Adenylate Kinase
AT1G62830	LOC_Os02g51880	cassava4.1_001812	chromatin complex modification, root elongation, repressor of floral transition gene (FT gene)	Lysine-Specific Histone Demethylase-1 Homolog 1
AT5G51820	LOC_Os10g11140	cassava4.1_003471	starch biosynthesis, trehalose biosynthesis, positive regulator of polysaccharide catabolism	Glucose-1-phosphate phosphotransferase
AT5G48300	LOC_Os08g25734	cassava4.1_005446	starch and glycogen biosynthesis, photoperiodism, flowering	Glucose-1-phosphate adenyl transferase small subunit chloroplastic
AT5G19220	LOC_Os03g52460	cassava4.1_028396	starch and glycogen biosynthesis, RNA processing	Glucose-1-phosphate adenyl transferase large subunit, chloroplastic

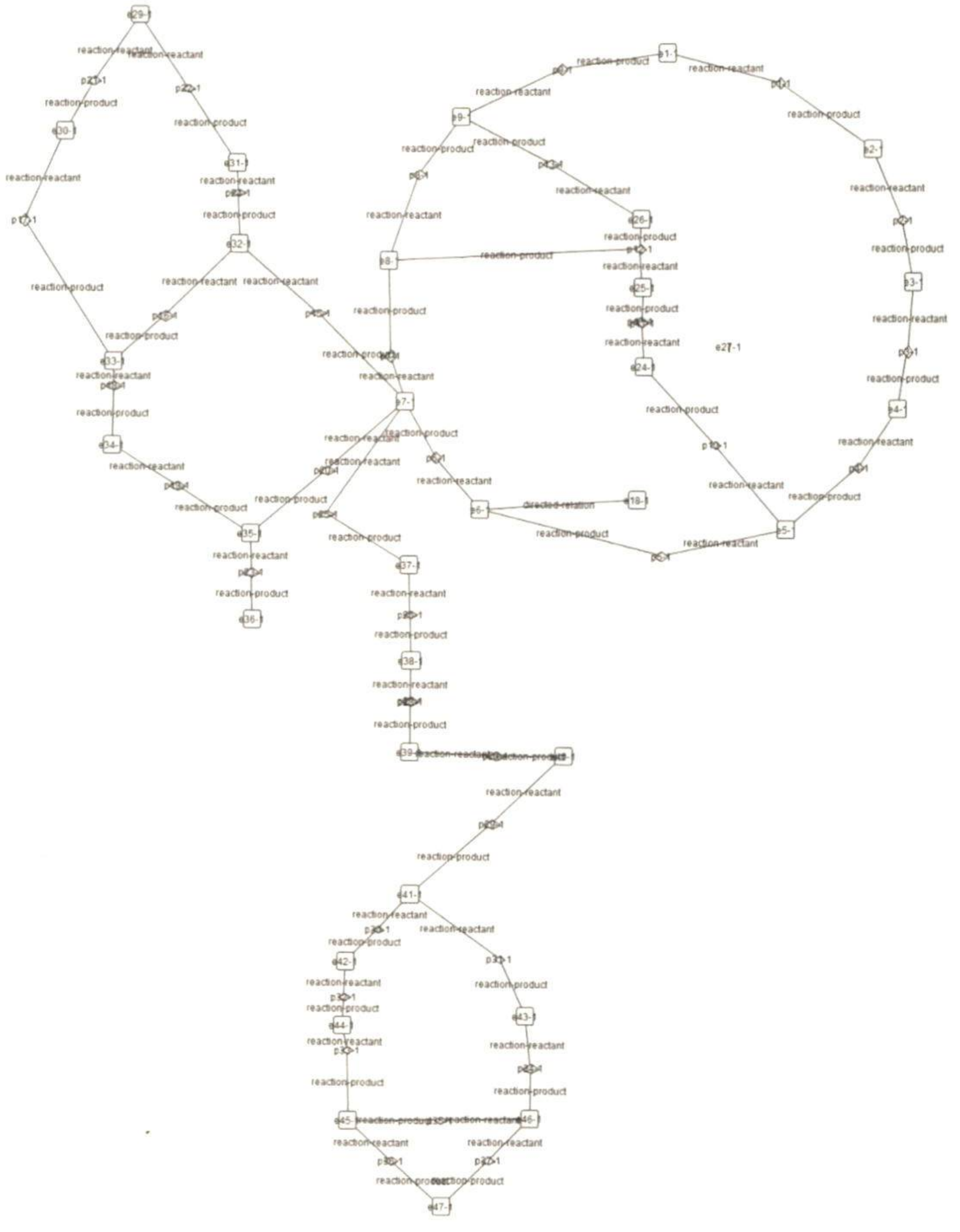

Fig 7. Consolidated pathway of starch biosynthesis in cassava viewed on VANTED tool

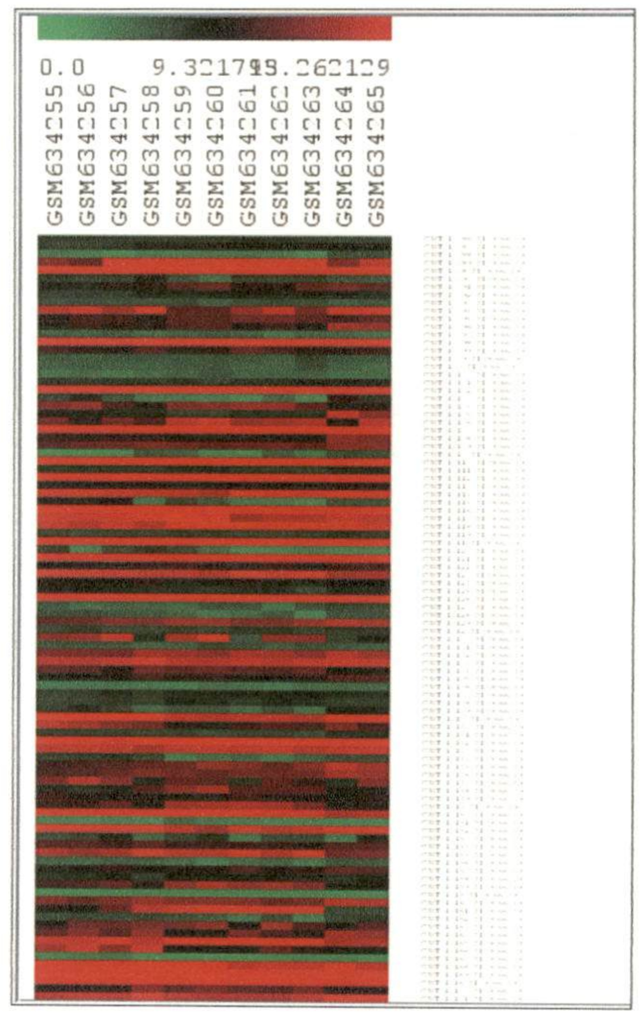

Fig 8. Heat map of the genes after normalization generated from GeneXplain platform

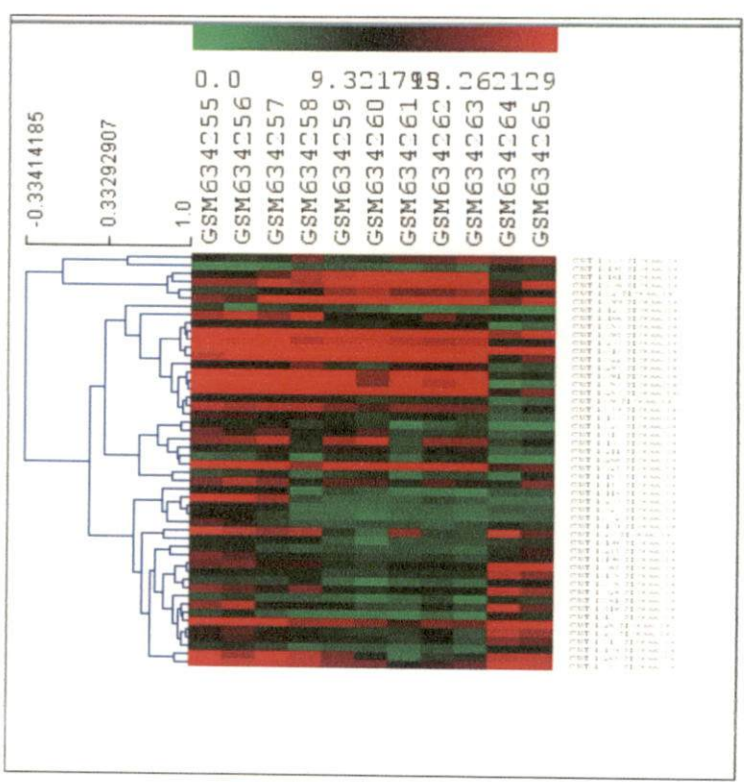

Fig 9. HCL clustering of the genes after normalization generated from MeV software

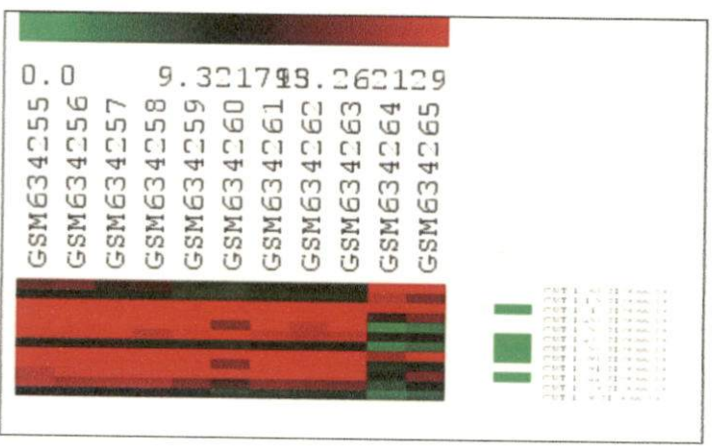

Fig 10. Heat map of the significant genes generated from MeV software

Table 12. Differentially expressed genes identified from microarray gene expression analysis

Differentially Expressed Genes with $SNR > 2.5$ and $P < 0.05$
Contig2017
TA9083_3983
TA5570_3983
Contig5561
TA8522_3983
Contig4441
BM259732

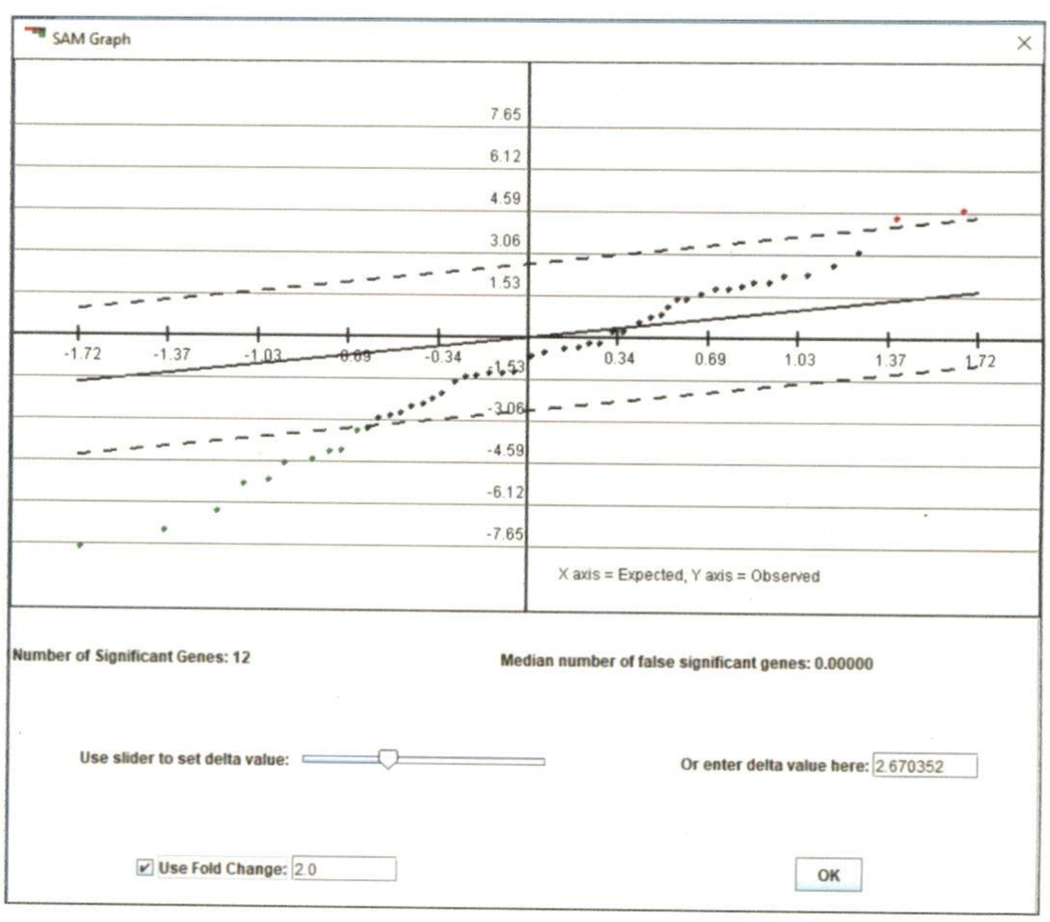

Fig 11. SAM graph of significant genes from the MeV software

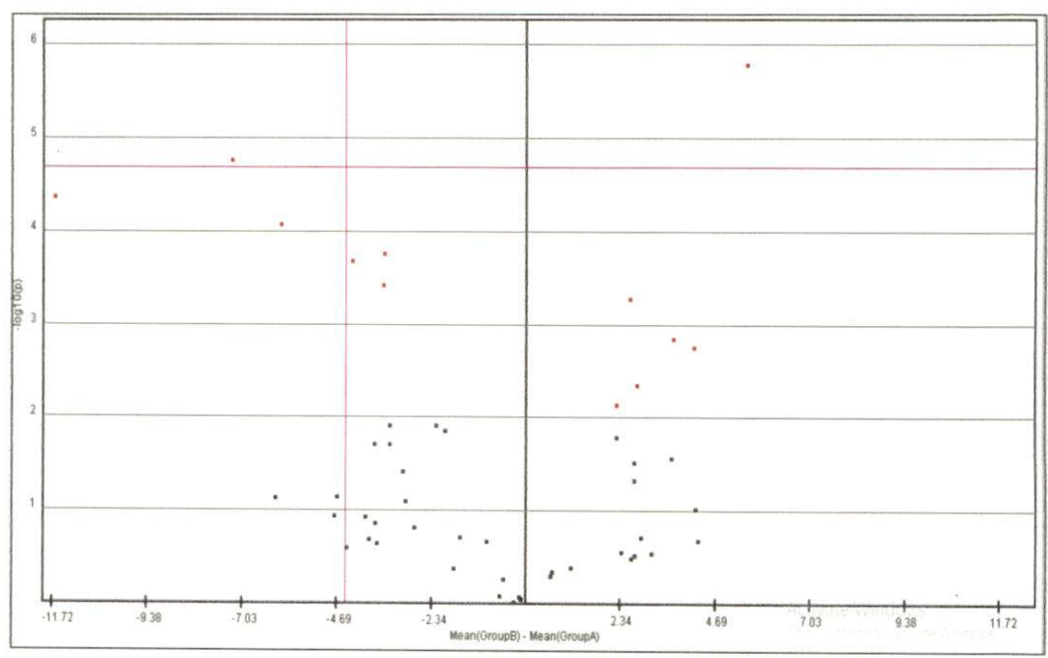

Fig 12. Volcano plot of the significant genes from MeV software

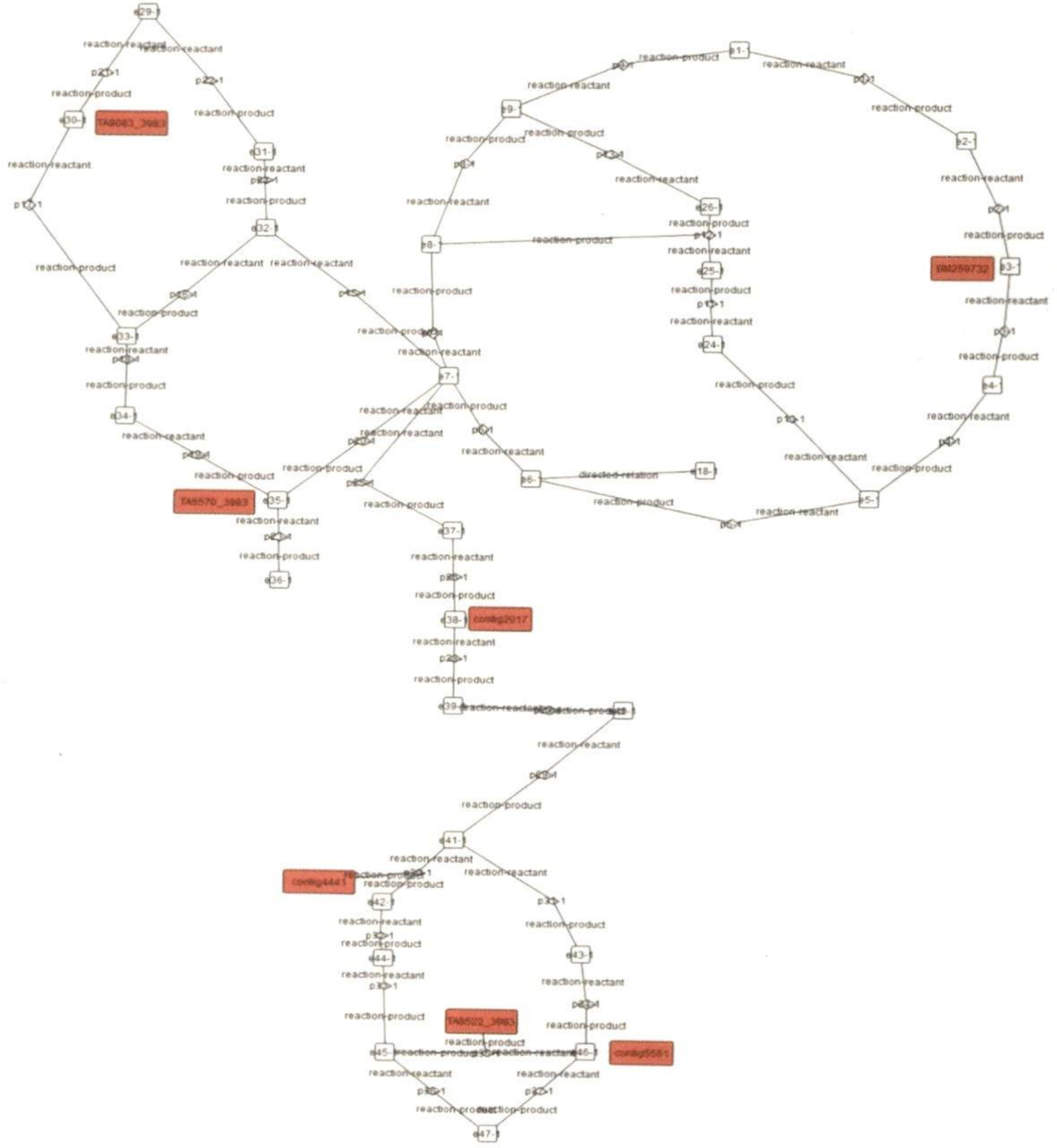

Fig 13. The interactive pathway of reconstructed starch biosynthesis pathway in cassava along with the integration of differentially expressed genes viewed on VANTED platform

DISCUSSION

5. DISCUSSION

The study entitled "Comparative and Functional Genomics Analysis of Starch Biosynthesis Pathways in Cassava" was conducted to investigate on the starch biosynthesis pathways in cassava using data integration and to identify the differentially expressed genes during different growth stages of cassava. The results of this study presented in chapter 4 are discussed here.

Cassava is an important food crop next to rice and maize and is the staple food crop to a large population around the world. Compared to other crops where starch is the carbohydrate reserve, cassava is unique in various aspects like the capacity for starch accumulation, being drought tolerant, the deprivation of soil nutrients and the utmost ability of the crop to grown versatile under water scarce and low fertilized soil. With an increase in the global population along with the strong demand for starch, most of the research focuses on cassava to produce novel cassava lines with greater starch yield per unit area.

In contrast with the relevance of cassava, scarce amount of information on cassava is presently available in the literature. Prior to 2009, very little percentage of the cassava genome was revealed which was mostly collected and disseminated from the classical genetic and EST expression studies (Prochnik *et al.*, 2012). During the earlier times, the cassava genetic information was exploited and that information was utilized to breed for a novel cassava line (El-Sharkawy, 2004). The present era is revolutionised with several modern as well as high throughput technologies to investigate the complex regulation underlying the biological processes in cassava (Sakurai *et al.*, 2007; Yang *et al.*, 2011). There are many innovations and technologies for inspecting the intracellular components such as the microarray and C-13 labelling experiment that enabled researchers to investigate the regulation controlling the behaviour of cells. With the release of the cassava genome sequence in 2009 the margin of the cassava research has leaped forward (Goodstein *et al.*, 2011). The present sequenced genome of cassava covers 69 percent of the predicted genome size and contains 96 percent of the whole generic region. All the 30,666 genes and 3,485 alternative splicing

forms were predicted and annotated in cassava. The entire genome information has been made available at the Phytozome database (Prochnik *et al.*, 2012). Even though, sequencing had shed light into the cassava genome information but the real complete genome annotation is on the progress. In addition to the relatively little background knowledge on cassava in plant science research, researchers and the plant breeders are highly interested in understanding complicated regulation underlying starch biosynthesis.

The present knowledge of starch biosynthesis in cassava relies very much on the observations in the Arabidopsis model plant and other starchy crops such as the potato. The plant species share the backbone pathway of starch biosynthesis in common starting from the carbon dioxide fixation or Calvin cycle succeeded by transitory starch degradation, sucrose synthesis, and starch synthesis in the storage organs. There is a clear cut distinction between the starch biosynthesis pathways of cereal and non-cereal crops which provides evidence of pathway diversity in plant species. The variation in starch biosynthesis in cereals and non-cereals involves the conversion of glucose-6-phosphate (G6P) to ADP-glucose (ADPG), a common precursor for starch synthesis. In cereals, the conversion takes place in the cytosol and through the amyloplast, ADPG is transported through a membrane-bound transporter, where as in non-cereal crops, G6P is directly imported to the amyloplast prior to the G6P-to-ADPG conversion (Comparot-Moss and Denyer, 2009). Thus, to pursue a better understanding of starch biosynthesis in cassava, the species-specific pathway demonstrating a series of reactions orchestrated in the network of the cassava starch production process is required.

The metabolic pathways which are species- specific have continuously been reconstructed not only for the broad purposes as found in the public data repositories but also for a particular pathway investigation as demonstrated in various publications. Apart from the model plant Arabidopsis (Poolman *et al.*, 2009; De oliveira *et al.*, 2010), the starch biosynthesis pathways have been reconstructed in potato (Poolman *et al.*, 2009), wheat (Dupont, 2008) and maize

(Friso *et al.*, 2010). In cassava, only broad metabolic pathways have been investigated (Sakurai *et al.*, 2007). There were several attempts to study and disseminate information about the starch biosynthesis process in cassava (Ihemere *et al.*, 2006). Sakurai *et al.* (2007) reconstructed the first metabolic pathways of cassava using the comparative genomic analysis of the full-length cDNA library even before the release of the cassava genome data. However, there were several pathway holes in the pathway reconstructed. Rongsirikul *et al.* (2010) modified and reconstructed the starch biosynthesis pathway in cassava employing inverse genomics approach shortly after the availability of the cassava genome sequence in the Phytozome database. Finally another reconstruction protocol was developed by Saithong *et al.* (2013) exploiting five template plants other than cassava namely Arabidopsis, rice, maize, castor-bean and potato. The parsimonious reconstruction protocol followed various intensive analyses to ensure high-level of quality of the resulting pathway.

The starch biosynthesis process was seen to be conserved in among the six plant species including cassava and the five template plants which was observed at both molecular sequence and pathway levels. In analogy to the orthologous or conserved gene identification from the similarity of sequences, the conservation at the pathway level may be inferred from the similarity of the constituents and the structure of the pathway. Starch-storage organs were physiologically distinct in most of the plants but the molecular sequence of various enzymes involved in the process remains evolutionarily conserved. For example, the genes encoding for sucrose synthase (EC 2.4.1.13) and ADP-glucose pyrophosphorylase (EC 2.7.7.27) which are the major constituents of starch biosynthesis were found in nearly all plant species and the orthologous sequences show a comparatively higher range of sequence similarity with the sequence of cassava (Zeeman *et al.*, 2010; Sonnewald and Kossmann, 2013). At the pathway level, the starch biosynthesis pathway of cassava showed resemblance with the pathways of starch synthesis in other plants retrieved from the KEGG and PMN reference databases in terms of the number of enzymatic proteins and biochemical pathways.

However, the reconstruction protocol by Rongsirikul *et al.* (2010) and Saithong *et al.* (2013) lacks molecular insights of pathway regulation of starch metabolism in cassava. Hence, another strategy called the gene prioritisation approach was carried out in the present study to identify the candidate genes involved in the pathways of starch synthesis in cassava.

Genotype-to phenotype relationship is crucial for the plant breeders to unravel the plant genomics data for developing elite traits. The molecular and biological basis of most quantitative trait variations is poorly understood in majority of plants. QTL mapping approaches have got several limitations such as the generation of large numbers of candidate genes which make it difficult to identify causal genes easily. The prioritization of candidate genes is not only of fundamental interest, but also of high practical value, because causal genes for any trait-of interest make perfect markers for breeding. 5 genes have been prioritised in the present study that can serve as candidate genes in the starch biosynthesis pathways in cassava. One of the four genes prioritized, cassava4.1_013232 with an annotated function as that of plastidial Adenylate kinase act as a negative regulator of starch synthesis (Oliver *et al.*, 2008). Geigenberger *et al.* (2004) suggested that plastidial adenylate kinase is a negative regulator of starch synthesis with a very strong negative control coefficient implying that it competes for ATP with starch and amino acid under normal physiological conditions.

The reconstructed pathway was finally used as a platform for transcriptome data integration to enhance our comprehension on the starch biosynthesis in cassava. Data integration enabled us to yield a better understanding into the starch biosynthesis process through investigation on the multilevel regulations and interrelations between them. To demonstrate graphically the benefit of our approach and the innovated pathway, the microarray gene expression data of Yang *et al.* (2011) were integrated into the cassava starch biosynthesis pathway in the VANTED platform and cross-validated the gene expression patterns during different growth stages. The results suggested the

174176

KERALA AGRICULTURAL UNIVERSITY
CENTRAL LIBRARY
K.A.U. (P.O.), THRISSUR

dominance of C3 or Calvin cycle over C4 pathway when the gene expression analysis was compared with C3 and C4 plants. Being a C3 C4 intermediate, in cassava, gene expression analysis revealed that the expression of genes in the Calvin cycle evolves with the root types where as the C4 pathway is tentatively independent of the types of cassava roots. Our results also showed that the activities of enzymes involved in the C4 pathway are less dominant compared to the enzymes that actively participate in the Calvin cycle (Edwards *et al.*, 1990). For the starch and sucrose biosynthesis pathways in cassava, the integrated map suggested a sensible demonstration of the transition in root development of cassava along with the progressive starch production. Finally, it was inferred that the gene expression of the enzymes that catalyse reactions toward starch production is highly expressed in the mature storage roots with respect to that in the fibrous and developing storage roots, indicating the dominance of this metabolic process in the storage root type.

The pathway was reconstructed with three putative enzyme genes and the transcription factors involved in the starch biosynthesis in cassava. The transcriptome data integration finally revealed that there are 7 genes that are differentially expressed in different processes of starch biosynthesis namely contig2017, TA9083_3983, TA5570_3983, Contig5561, TA8522_3983, Contig4441 and BM259732. The new reconstructed pathway could be used as a better tool by the plant breeders in near future for the development of novel cassava cultivars with higher starch yield.

SUMMARY

6. SUMMARY

The study entitled "Comparative and Functional Genomics Analysis of Starch Biosynthesis Pathways in Cassava" was carried out at the Section of Extension and Social Sciences, ICAR-Central Tuber Crops Research Institute, Sreekariyam, Thiruvananthapuram during 2016-2017. The objectives of the study were to study on the starch biosynthesis pathways in cassava using data integration and temporal gene expression pattern during different growth stages.

The study was divided into two phases: pathway reconstruction and transcriptome data integration. The starch biosynthesis pathway in cassava comprised of carbon-dioxide fixation, sucrose biosynthesis I (from photosynthesis), sucrose biosynthesis II and starch biosynthesis finally yielding starch, which is the major carbohydrate reservoir of the plants. BLAST template preparation was the initial step wherein the starch synthesising genes and proteins were collected from multiple template plants namely arabidopsis, maize, rice castor bean and potato. BLASTp and reciprocal BLASTp were carried out simultaneously to identify the best reciprocal BLAST hits (RBHs). Three enzyme putative gene sequences were obtained which were translated into protein sequences using TranSeq, a tool under EBI. Further regulatory sequence analysis and protein motif analysis were carried out. Protein motif analysis was carried out using The MEME Suite which contains various tools for motif analyses *viz*., MEME, MAST, TOMTOM and GOMO. Protein- protein interaction networks were also created using STITCH and graphical output was obtained. Regulatory sequence analyses primarily include promoter analysis via ProScan and transcription factor binding site predictions using PlantTFDB (Plant Transcription Factor Database), Alggen-PROMO and JASPAR. Putative ten transcription factor binding sites and also TF signals in the promoter sequences were obtained after the analysis.

The core part of the work involved is the pathway reconstruction of starch biosynthesis in cassava. Prior to it, another approach named "Gene Prioritisation" was employed in the study. Through this strategy, potential candidate genes involved in starch synthesis could be identified which is thought to revolutionise the field of plant breeding. Gene prioritisation tools used in the study were ToppGene suite and RiceNet. ToppGene suite involves tools for gene prioritisation and prioritising genes in protein-protein interaction. RiceNet is a gene prioritisation tool that is exclusively used for the identification of candidate genes in rice. Using this approach, five genes were prioritised of which one is a negative regulator of starch synthesis. The negative regulator identified codes for the enzyme plastidial adenylate kinase whose role in starch synthesis was studied recently by Oliver *et al.* (2008).

Final module of the work was pathway reconstruction and microarray data integration. The construction of a consolidated pathway of starch biosynthesis process was carried out in Cell Illustrator. Regulatory elements like transcription factors and candidate genes were done through the Cell Illustrator software. The reconstructed pathway was viewed on VANTED (Visualisation and Analysis of Networks containing Experimental Data) platform. To get a better comprehension of the reconstructed pathway, transcriptome data of the TMS60444 cultivar of cassava with dataset GSM25813 was collected form NCBI GEO. Three treatments with replicates namely developing root, fibrous root and mature root along with controls (storage root slice) were taken for the identification of differentially expressed genes (DEGs). The controls taken were used as the technical repeats in the study. From a large dataset, 5000 genes were taken for the study. After normalisation, it was found that 500 genes participate in the starch synthesis. Applying a statistical cut off of SNR ratio greater than 2.58 and P-value less than 0.05, the number of genes were again reduced to 48. All the statistical analyses were performed on the GeneXplain platform. In the MeV (Multi Experiment Viewer), Heat map was constructed, identified for the enumeration of significant genes. 12 significant genes were obtained which were thought to take

part in the starch synthesis pathway. Statistical analysis was carried out using Hierarchial clustering (HCL), Significance analysis of Microarray (SAM) graph and Volcano Plot. Seven differentially expressed genes were obtained namely Contig2017, TA9083_3983, TA5570_3983, Contig5561, TA8522_3983, Contig4441 and BM259732. The differentially expressed genes were seen to express more in the storage roots than in the developing and fibrous roots. Finally, a complete pathway of starch synthesis was recreated after the incorporation of regulatory elements, candidate genes and expression data of differentially expressed genes (DEGs) and viewed on VANTED v2.6.3.

The reconstructed pathway of starch synthesis in cassava can be exploited in cassava breeding programs and has a high potential in understanding the comprehension behind cassava starch metabolism. Also the reconstructed pathway is a better tool that nullifies the pathway holes that were present in the previous reconstructed pathways of cassava.

REFERENCES

7. REFERENCES

Akihiro, T., Mizuno, K., and Fujimura, T. 2005. Gene expression of ADP-glucose pyrophosphorylase and starch contents in rice cultured cells are cooperatively regulated by sucrose and ABA. *Plant Cell Physiol.* 46: 937-946.

Alexander, R. J. 1995. Starch in plastics. *Cereal Foods World*. 41: 426-427.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215(3): 403-410.

Anderson, J. V., Delseny, M., Fregene, M. A., Jorge, V., Mba, C., Lopez, C., Restrepo, S., Soto, M., Piegu, B., Verdier, V., Cooke, R., Tohme, J., and Horvath, D. P. 2004. An EST resource for cassava and other species of Euphorbiaceae. *Plant Mol. Biol.* 56: 527-539.

Anthony, M., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Zhou, M., and Wasserman, W. W. 2013. JASPAR 2013: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 42: 142-147.

Arijaje, E. O., Jayanthi, S., Kumar, T. K. S., and Wang, Y. J. 2015. Linear starch and hexanoic acid complexation evaluated by isothermal titration calorimetry. *Starch/Staerke*. 67: 729-736.

Backiyarani, S., Uma, S., Saraswathi, M. S., Saravanakumar, A. S., and Chandrasekar, A. 2015. Transcriptome analysis of banana (Musa balbisiana) based on next-generation sequencing technology. *Turk. J. Agric. For.* 39: 705-717.

Baguma, Y. 2004. Regulation of Starch Synthesis in Cassava. Doctoral thesis, Swedish University of Agricultural Sciences, Uppsala, 35p.

Balat, M. and Balat, H. 2009. Recent trends in global production and utilization of bio-ethanol fuel. *Appl. Energy*. 86: 2273–2282.

Ball, S. G. and Morell, M. K. 2003. From bacterial glycogen to starch: Understanding the Biogenesis of the Plant Starch Granule. *Annu. Rev. Plant Biol.* 54: 207–233.

Baroja-Fernández, E., Muñoz, F.J., Montero, M., Etxeberria, E., Sesma, M. T., Ovecka, M., Bahaji, A., Ezquer, I., Li, J., Prat, S., and Pozueta-Romero, J. 2009. Enhancing sucrose synthase activity in transgenic potato (Solanum tuberosum L.) tubers results in increased levels of starch, ADPglucose and UDPglucose and total yield. *Plant cell physiol.* 50: 1651-1662.

Beckles, D. M. 2001. ADP-Glucose Pyrophosphorylase Is Located in the Plastid in Developing Tomato Fruit. *Plant Physiol.* 126: 261-266.

Benesi, I., Labuschagne, M., Dixon, A., and Mahungu, N. 2004. Genotype x environment interaction effects on native cassava starch quality and potential for starch use in the commercial sector. *Afr. Crop Sci. J.* 12(3): 205-216.

Blesak, K. and Janecek, T. 2012. Sequence fingerprints of enzyme specificities from the glycoside hydrolase family GH57. *Extremophiles*.16: 497-506.

Breton, C., Šnajdrová, L., Jeanneau, C., Koča, J., and Imberty, A. 2006. Structures and mechanisms of glycosyltransferases. *J. Glycobiol.* 16: 29-37.

Brummell, D. A., Watson, L. M., Zhou, J., McKenzie, M. J., and Hallett, I. C. 2015. Over-expression of starch branching enzyme II increases short-chain branching of amylopectin and alters the physicochemical properties of starch from potato tuber. *BMC Biotechnol.* 15: 28.

Buleon, A., Colonna, P., Planchot, V., and Ball, S. 1998. Starch granules: structure and biosynthesis. *Int. J. Biol. Macromol.* 23(2): 85-112.

Burlingame, B., Charrondiere, R., and Mouille, B. 2009. Food composition is fundamental to the cross-cutting initiative on biodiversity for food and nutrition. *J. Food Compos. Anal.* 22: 361-365.

Burns, A. E., Gleadow, J., Cliff, J., Zacarias, A., and Cavagnaro, T. 2010. Cassava: the drought, war and famine crop in a changing world. *Sustain.* 2: 3572–3607.

Burton, R. A., Bewley, J. D., Smith, A. M., Bhattacharyya, M. K., Tatge, H., Ring, S., Bull, V., Hamilton, W.D., and Martin, C. 1995. Starch branching enzymes belonging to distinct enzyme families are differentially expressed during pea embryo development. *Plant J.* 7(1): 3-15.

Bustos, R., Fahy, B., Hylton, C. M., Seale, R., Nebane, N. M., Edwards, A., Martin, C., and Smith, A. M. 2004. Starch granule initiation is controlled by a heteromultimeric isoamylase in potato tubers. *Proc. Natl. Acad. Sci.* 101: 2215-2220.

Carciofi, M., Blennow, A., Jensen, S. L., Shaik, S. S., Henriksen, A., Buleon, A., Holm, P. B., and Hebelstrup, K. H. 2012. Concerted suppression of all starch branching enzyme genes in barley produces amylose-only starch granules. *BMC Plant Biol.* 12: 223.

Ceballos, H., Hershey, C., and Becerra-López-Lavalle, L. A. 2012. New approaches to cassava breeding. *Plant Breed. Rev.* 36: 427-504.

Ceballos, H., Iglesias, C. A., Pérez, J. C., and Dixon, A. G. 2004. Cassava breeding: opportunities and challenges. *Plant Mol Biol.* 56(4): 503-516.

Chae, L., Kim, T., Nilo-Poyanco, R., Rhee, S. Y. 2014. Genomic signatures of specialized metabolism in plants. *Sci.* 344: 510-513.

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37: 305-311.

Chen, R., Mallelwar, R., Thosar, A., Venkatasubrahmanyam, S., and Butte, A. J. 2008.GeneChaser: identifying all biological and clinical conditions in which genes of interest are differentially expressed. *BMC bioinforma.* 9: 548.

Chetty, C. C., Rossin, C. B., Gruissem, W., Vanderschuren, H., and Rey, M. E. C. 2013. Empowering biotechnology in southern Africa: Establishment of a robust transformation platform for the production of transgenic industry-preferred cassava. *New Biotechnol.* 30: 136-143.

Colleoni, C., Steven, G., and Suzuki, E. 2017.Structural basis of substrate binding of branching enzyme. *J. Biol. Chem.* 292: 1-29.

Comparot-Moss, S., and Denyer, K. 2009. The evolution of the starch biosynthetic pathway in cereals and other grasses. *J. Exp. Bot.* 60: 2481-2492.

Craig, J., Lloyd, J. R., Tomlinson, K., Barber, L., Edwards, A., Wang, T. L., Martin, C., Hedley, C. L., and Smith, A. M. 1998. Mutations in the gene encoding Starch Synthase II profoundly alter amylopectin structure in pea embryos. *Plant Cell.* 10: 413-426.

Crevillen, P., Ballicora, M. A., Merida, A., Preiss, J., and Romero, J. M. 2003. The different large subunit isoforms of Arabidopsis thaliana ADP-glucose pyrophosphorylase confer distinct kinetic and regulatory properties to the heterotetrameric enzyme. *J. Biol. Chem.* 278: 28508-28515.

Crevillen, P., Ventriglia, T., Pinto, F., Orea, A., Mérida, Á., and Romero, J. M. 2005. Differential pattern of expression and sugar regulation of Arabidopsis thaliana ADP-glucose pyrophosphorylase-encoding genes. *J. Biol. Chem.* 280: 8143-8149.

Da, G., Dufour, D., Giraldo, A., Moreno, M., and Tran, T. 2013. Cottage level cassava starch processing systems in Colombia and Vietnam. *Food Bioprocess Technol.* 6(8): 2213-2222.

de Oliveira Dal'Molin, C. G., Quek, L. E., Palfreyman, R.W., Brumbley, S. M., and Nielsen, L. K. 2010. AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis. *Plant Physiol.* 152: 579-589.

Delatte, T., Trevisan, M., Parker, M. L., and Zeeman, S. C. 2005. Arabidopsis mutants Atisa1 and Atisa2 have identical phenotypes and lack the same multimeric isoamylase, which influences the branch point distribution of amylopectin during starch synthesis. *Plant J.* 41: 815-830.

Delvalle, D., Dumez, S., Wattebled, F., Roldán, I., Planchot, V., Berbezy, P., Colonna, P., Vyas, D., Chatterjee, M., Ball, S., Mérida, A., and D'Hulst, C. 2005. Soluble starch synthase I: A major determinant for the synthesis of amylopectin in Arabidopsis thaliana leaves. *Plant J.* 43: 398-412.

Denyer, K., Dunlap, F., Thorbjornsen, T., Keeling, P., and Smith, A. M. 1996.The Major Form of ADP-Glucose Pyrophosphorylase in Maize Endosperm 1s Extra-Plastidial. *Plant Physiol.* 112: 779-785.

Denyer, K., Dunlap, F., Thorbjornsen, T., Keeling, P., and Smith, A. M. 1996. The Major Form of ADP-Glucose Pyrophosphorylase in Maize Endosperm 1s Extra-Plastidial. *Plant Physiol.* 112: 779-785.

Dinges, J. R., Colleoni, C., James, M. G., and Myers, A. M. 2003. Mutational analysis of the pullulanase-type debranching enzyme of maize indicates multiple functions in starch metabolism. *Plant cell.* 15: 666-680.

Djerbi, S., Lindskog, M., Arvestad, L., Sterky, F., and Teeri, T. T. 2005. The genome sequence of black cottonwood (Populus trichocarpa) reveals 18 conserved cellulose synthase (CesA) genes. *Planta.* 221: 739-746.

Domań-Pytka, M. and Bardowski, J. 2004. Pullulan degrading enzymes of bacterial origin. *Crit. Rev. Microbiol.* 30: 107-121.

Dumez, S., Wattebled, F., Dauvillee, D., Delvalle, D., Planchot, V., Ball, S. G., and D'hulst, C. 2006. Mutants of Arabidopsis lacking starch branching enzyme II substitute plastidial starch synthesis by cytoplasmic maltose accumulation. *Plant Cell.* 18: 2694-2709.

Dupont, F. M. 2008. Metabolic pathways of the wheat (Triticum aestivum) endosperm amyloplast revealed by proteomics. *BMC plant biol.* 8: 39.

Edgar, R., Domrachev, M., and Lash, A. E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30(1): 207-210.

Edwards, A., Fulton, D. C., Hylton, C. M., Jobling, S. A., Gidley, M., Rössner, U., Martin, C., and Smith, A. M. 1999. A combined reduction in activity of starch synthases II and III of potato has novel effects on the starch of tubers. *Plant J.* 17: 251-261.

Edwards, A., Vincken, J. P., Suurs, L. C. J. M., Visser, R. G. F., Zeeman, S., Smith, A., and Martin, C. 2002. Discrete forms of amylose are synthesized by isoforms of GBSSI in pea. *Plant cell.* 14: 1767-1785.

Edwards, J. W., Walker, E. L., and Coruzzi, G. M. 1990. Cell-specific expression in transgenic plants reveals non overlapping roles for chloroplast and cytosolic glutamine synthetase. *Proc. Natl. Acad. Sci.* 87: 3459-3463.

El-Sharkawy, M. A. 2006. International research on cassava photosynthesis, productivity, eco-physiology, and responses to environmental stresses in the tropics. *Photosynthetica.* 44(4): 481-512.

Emes, M. J., Bowsher, C. G., Hedley, C., Burrell, M. M., Scrase-Field, E. S. F., and Tetlow, I. J. 2003. Starch synthesis and carbon partitioning in developing endosperm. *J. Exp. Bot.* 54: 569-575.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nat.* 402: 86-90.

Espada, J. 1962. Enzymatic Synthesis of Adenosine Diphosphate Glucose from Glucose 1-Phosphate and Adenosine Triphosphate. *J. Biol. Chem.* 237: 3577-3581.

Fabian, A. B., Mikael, B., Denis, C. B., and Timothy, L. B. 2010. Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinforma.* 26(7): 860-866.

FAO (1990) Food and Agricultural Organization of the United Nations, Rome. http://www.fao.org/ag/agp/agpc/gcds/index_en.html. [22 Dec. 2016].

FAO (2008) Food and Agricultural Organization of the United Nations, Rome. http://www.fao.org/ag/agp/agpc/gcds/index_en.html. [22 Dec. 2016].

FAO (2013) Food and Agricultural Organization of the United Nations, Rome. http://www.fao.org/ag/agp/agpc/gcds/index_en.html. [22 Dec. 2016].

FAOSTAT (2013) Food and agriculture organizations statistics database. FAO, Rome. http://www.fao.org/faostat/en.html. [10 Jan. 2017].

FAOSTAT (2017) Food and agriculture organizations statistics database. FAO, Rome. http://www.fao.org/faostat/en.html. [07 Jan. 2017].

Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., Yan, A., Mueller, L. A. 2015. The Sol Genomics Network (SGN) – from genotype to phenotype to breeding. *Nucleic Acids Res*. 43: 1036-1041.

Fisher, D. K., Gao, M., Kim, K. N., Boyer, C. D., and Guiltinan, M. J. 1996. Two closely related cDNAs encoding starch branching enzyme from Arabidopsis thaliana. *Plant Mol. Biol*. 30: 97-108.

Frehner, M., Pozueta-Romero, J., and Akazawa, T. 1990. Enzyme Sets of Glycolysis, Gluconeogenesis, and Oxidative Pentose Phosphate Pathway Are Not Complete in Nongreen Highly Purified Amyloplasts of Sycamore (Acer pseudoplatanus L.) Cell Suspension Cultures. *Plant physiol.* 94: 538-544.

Friso, G., Majeran, W., Huang, M., Sun, Q., and Wijk, K. J. 2010. Reconstruction of metabolic pathways, protein expression, and homeostasis machineries across maize bundle sheath and mesophyll chloroplasts: large-scale quantitative proteomics using the first maize genome assembly. *Plant Physiol*. 152: 1219-1250.

Fujita, N., Toyosawa, Y., Utsumi, Y., Higuchi, T., Hanashiro, I., Ikegami, A., Akuzawa, S., Yoshida, M., Mori, A., Inomata, K., Itoh, R., Miyao, A., Hirochika, H., Satoh, H., and Nakamura, Y. 2009. Characterization of pullulanase (PUL)-deficient mutants of rice (Oryza sativa L.) and the

function of PUL on starch biosynthesis in the developing rice endosperm. *J. Exp. Bot.* 60: 1009-1023.

Gamm, M., Heloir, M. C., Bligny, R., Vaillant-Gaveau, N., Trouvelot, S., Alcaraz, G., Frettinger, P., Clement, C., Pugin, A., Wendehenne, D., and Adrian, M. 2011. Changes in carbohydrate metabolism in Plasmopara viticola-infected grapevine leaves. *Mol. Plant-Microbe Interact.* 24: 1061–1073.

Geigenberger, P., Stamme, C., Tjaden, J., Schulz, A., Quick, P.W., Betsche, T., Kersting, H. J., and Neuhaus, H. E. 2001. Tuber physiology and properties of starch from tubers of transgenic potato plants with altered plastidic adenylate transporter activity. *Plant Physiol.* 125: 1667-1678.

Geigenberger, P., Stitt, M., and Fernie, A. R. 2004. Metabolic control analysis and regulation of the conversion of sucrose to starch in growing potato tubers. *Plant Cell Environ.* 27: 655-673.

Gelfand, M. S. 1999. Recognition of regulatory sites by genomic comparison. *Res. Microbiol.* 150: 755-771.

Georgelis, N., Braun, E. L., and Hannah, L. C. 2008. Duplications and functional divergence of ADP-glucose pyrophosphorylase genes in plants. *BMC Evol. Biol.* 8: 232.

Gerard, C., Colonna, P., Buleon, A., and Planchot, V. 2002. Order in maize mutant starches revealed by mild acid hydrolysis. *Carbohyd. Polym.* 48: 131-141.

Gibson, K., Park, J. S., Nagai, Y., Hwang, S. K., Cho, Y. C. 2011. Exploiting leaf starch synthesis as a transient sink to evaluate photosynthesis, plant productivity and yields. *Plant Sci.* 181: 275-281.

Giroux, M. J. and Hannah, L. C. 1994.ADP-glucose pyrophosphorylase in shrunken-2 and brittle-2 mutants of maize. *Mol. Gen. Genet.* 243: 400–408.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., and Putnam, N. 2011. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40: 1178-1186.

Hannah, L. C. 1997. Starch synthesis in the maize seed. *Cell. Mol. Biol. Plant Seed Dev.* 4: 375-405.

Hatzfeld, W. D. and Stitt, M. 1990. A study of the rate of recycling of triose phosphates in heterotrophic Chenopodium rubrum cells, potato tubers, and maize endosperm. *Planta*. 180: 198-204.

Hawker, J. S., Marschner, H., and Krauss, A. 1979. Starch Synthesis in Developing Potato-Tubers. *Physiologia Plant.* 46: 25-30.

Hayat, I., Ahmad, A., Masud, T., Ahmed, A., and Bashir, S. 2014. Nutritional and health perspectives of beans (Phaseolus vulgaris L.): an overview. *Crit. Rev. Food Sci. Nutr.* 54: 580-592.

Hebelstrup, K. H., Christiansen, M.W., Carciofi, M., Tauris, B., Brinch-Pedersen, H., and Holm, P. B. 2010. UCE: A uracil excision (USER)-based toolbox for transformation of cereals. *Plant methods.* 6: 15.

Hendrik, R., Astrid, J., Anja, H., Eva, G., Hendrik, T., and Christian, T. 2012. VANTED v2: a framework for systems biology applications. *BMC Syst. Biol.* 6: 139.

Hu, X., Kuang, S., Zhang, A., Zhang, W. S., Chen, M. J., Yin, X. R., and Chen, K. S. 2016. Characterization of starch degradation related genes in postharvest kiwifruit. *Int. J. Mol. Sci.* 17: 1-9.

Hussain, H., Mant, A., Seale, R., Zeeman, S., Hinchliffe, E., Edwards, A., Hylton, C., Bornemann, S., Smith, A. M., Martin, C., and Bustos, R. 2003. Three isoforms of isoamylase contribute different catalytic properties for the debranching of potato glucans. *Plant cell.* 15: 133-149.

Ihemere, U., Arias-Garzon, D., Lawrence, S., and Sayre, R. 2006. Genetic modification of cassava for enhanced starch production. *Plant Biotechnol. J.* 4: 453-465.

Ikeo, K., Ishi, J., Tamura, T., Gojobori, T., and Tateno, Y. 2003. CIBEX: center for information biology gene expression database. *Comptes Rendus Biologies.* 326(10): 1079-1082.

James, M. G., Robertson, D. S., Myers, A. M., James, M. G., Robertson, D. S., and Myers, A. N. 2015. Characterization of the Maize Gene sugary l, a Determinant Characterization in Kernels of Starch Composition. *Plant Cell.* 7: 417-429.

Jeon, J. S., Ryoo, N., Hahn, T. R., Walia, H., and Nakamura, Y. 2010. Starch biosynthesis in cereal endosperm. *Plant Physiol. Biochem.* 48: 383-392.

Jimenez, A., Fabra, M. J., Talens, P., andChiralt, A. 2012. Edible and Biodegradable Starch Films: A Review. *Food Bioprocess Technol.* 5: 2058-2076.

Jin, J. P., Zhang, H., Kong, L., Gao, G., and Luo, J. C. 2014. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res.* 42(1): 1182-1187.

Jobling, S. 2004. Improving starch for food and industrial applications. *Curr. Opin. Plant Biol.* 7(2): 210-218.

Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28: 27-30.

Karlstrom, A., Calle, F., Salazar, S., Morante, N., Dufour, D., and Ceballos, H. 2016. Biological Implications in Cassava for the Production of Amylose-Free Starch: Impact on Root Yield and Related Traits. *Front. Plant Sci.* 7: 604.

Kim, J. K., Park, S.Y., Kim, H. S., Kim, S. H., and Cho, K. S. 2005. Comparison of virtual cystoscopy, multiplanar reformation, and source CT images with contrast material-filled bladder for detecting lesions. *Am. J. Roentgenol.* 185: 689-696.

Knight, M. E., Harn, C., Lilley, C. E. R., Guan, H., Singletary, G.W., Muforster, C., Wasserman, B. P., and Keeling, P. L. 1998. Molecular cloning of starch synthase I from maize (W64) endosperm and expression in Escherichia coli. *Plant J.* 14: 613-622.

Kossmann, J., Abel, G. J., Springer, F., Lloyd, J. R., and Willmitzer, L. 1999. Cloning and functional analysis of a cDNA encoding a starch synthase from potato (Solanum tuberosum L.) that is predominantly expressed in leaf tissue. *Planta*. 6: 503-511.

Kubo, A., Colleoni, C., Dinges, J. R., Lin, Q., Lappe, R. R., Rivenbark, J. G., Meyer, A. J., Ball, S. G., James, M. G., Hennen-Bierwagen, T. A., and Myers, A. M. 2010. Functions of heteromeric and homomeric isoamylase-type starch-debranching enzymes in developing maize endosperm. *Plant Physiol.* 153: 956-969.

Kuhn, M., Damian, S., Thomas, H. B., Christian, V. M., Lars, J. J., and Peer, B. 2014. STITCH 4: integration of protein–chemical interactions with user data. *Nucleic Acids Res*. 42: 401-407.

Kunkeaw, S., Yoocha, T., Sraphet, S., Boonchanawiwat, A., and Boonseng, O. 2011. Construction of a genetic linkage map using simple sequence repeat markers from expressed sequence tags for cassava (Manihot esculenta Crantz). *Mol. Breed.* 27(1): 67-75.

Kwak, M. S., Min, S. R., Lee, S. M., Kim, K. N., Liu, J. R., Paek, K. H., Shin, J. S., andBae, J. M. 2007. A sepal-expressed ADP-glucose pyrophosphorylase gene (NtAGP) is required for petal expansion growth in 'Xanthi' tobacco. *Plant physiol.* 145: 277-289.

Lee, S. M., Ryu, T. H., Kim, S., Okita, T., and Kim, D. 2009. Kinetic and regulatory properties of plant ADP-glucose pyrophosphorylase genetically modified by heterologous expression of potato *upreg* mutants *in vitro* and *in vivo*. *Plant Cell Tissue Organ Cult.* 96: 161-170.

Lee, Tak, O., Taeyun , Y., Sunmo, S., Junha, W., Hwang, K., Yeong, K., Chan, K., Pam, L., 2015. RiceNet v2: an improved network prioritization server for rice genes. *Nucleic Acids Res*. 43: 122-127.

Leterrier, M., Holappa, L. D., Broglie, K. E., and Beckles, D. M. 2008. Cloning, characterisation and comparative analysis of a starch synthase IV gene in wheat: functional and evolutionary implications. *BMC Plant Biol*. 8: 98.

Li, J., Baroja-Fernandez, E., Bahaji, A., Munoz, F. J., Ovecka, M., Montero, M., Sesma, M.T., Alonso-Casajus, N., Almagro, G., Sanchez-Lopez, A. M., Hidalgo, M., Zamarbide, M., and Pozueta-Romero, J. 2013. Enhancing sucrose synthase activity results in increased levels of starch and ADP-glucose in maize (zea mays l.) seed endosperms. *Plant Cell Physiol.* 54: 282-294.

Lin, Q., Facon, M., Putaux, J. L., Dinges, J. R., Wattebled, F., D'Hulst, C., Hennen-Bierwagen, T. A., and Myers, A. M. 2013. Function of

isoamylase-type starch debranching enzymes ISA1 and ISA2 in the Zea mays leaf. *New Phytol.* 200: 1009-1021.

Lin, Q., Huang, B., Zhang, M., Zhang, X., Rivenbark, J., Lappe, R. L., James, M. G., Myers, A. M. & Hennen-Bierwagen, T. A. 2012. Functional interactions between starch synthase III and isoamylase-type starch-debranching enzyme in maize endosperm. *Plant Physiol.* 158: 679-692.

Lombard, V., Golaconda-Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42: 490-495.

Majoul-Haddad, T., Bancel, E., Martre, P., Triboi, E., and Branlard, G. 2013. Effect of short heat shocks applied during grain development on wheat (Triticum aestivum L.) grain proteome. *J. Cereal Sci.* 57: 486-495.

Makarova, K.S., Aravind, L., Grishin, N. V, Rogozin, I.B., and Koonin, E. V. 2002. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res.* 30: 482-496.

Manson, M. and Church, G. M. 2000. Conservation of DNA Regulatory Motifs and Discovery of New Motifs in Microbial Genomes Conservation of DNA Regulatory Motifs and Discovery of New Motifs in Microbial Genomes. *Genome Res.* 10: 744-757.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D.W., Yeates, T. O., Eisenberg, D., Fields, S., and Eisenberg, D. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Sci.* 285: 751-753.

Martin, C., and Smith, A. 1995. Starch biosynthesis. *Plant Cell.* 7(7): 971-985.

Meyer, F. D., Talbert, L. E., Martin, J. M., Lanning, S. P., Greene, T. W., and Giroux, M. J. 2007. Field evaluation of transgenic wheat expressing a

modified ADP-glucose pyrophosphorylase large subunit. *Crop Sci.* 47: 336-342.

Monaco, M. K., Stein, J., Naithani, S., Wei, S., Dharmawardhana, P., Kumari, S., Amarasinghe, V., Youens-Clark, K., Thomason, J., and Preece, J. 2014. Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res.* 42: 1193–1199.

Montagnac, J. A., Davis, C. R., and Tanumihardjo, S. A. 2009. Nutritional value of cassava for use as staple food and recent advances for improvement. *Compr. Rev. Food Sci. Food Saf.* 8(3): 181-194.

Morris, D. Z., and Morris, C. T. 1939. Glycogen in the seed of Zea mays (variety Golden Bantam). *J. Biol. Chem.* 130: 535-544.

Myers, A. M., Morell, M. K., James, M. G., and Ball, S. G. 2000. Recent progress toward understanding biosynthesis of the amylopectin crystal. *Plant physiol.* 122: 989-997.

Nagasaki, M., Saito, A., Jeong, E., Li, C., Kojima, K., Ikeda, E., and Miyano, S. 2010. Cell Illustrator 4.0: A computational platform for systems biology. *In Silico Biol.* 34: 45-56.

Nakamura, Y., Utsumi, Y., Sawada, T., Aihara, S., Utsumi, C., Yoshida, M., and Kitamura, S. 2010. Characterization of the reactions of starch branching enzymes from rice endosperm. *Plant Cell Physiol.* 51: 776-794.

Nardozza, S., Boldingh, H. L., Osorio, S., Höhne, M., Wohlers, M., Gleave, A. P., MacRae, E. A., Richardson, A. C., Atkinson, R.G., Sulpice, R., Fernie, A. R., and Clearwater, M. J. 2013. Metabolic analysis of kiwifruit (Actinidia deliciosa) berries from extreme genotypes reveals hallmarks for fruit starch metabolism. *J. Exp. Bot.* 64: 5049-5063.

Nelson, O. And Pan, D. 1995. Starch synthesis in maize endosperms. *Plant Mol. Biol*. 46: 475-496.

Neuhaus, H. E. and Wagner, R. 2000. Solute pores, ion channels, and metabolite transporters in the outer and inner envelope membranes of higher plant plastids. *Biochim. Biophys. Acta Biomembr*. 1465: 307-323.

Nighojkar, A. S. and Kumar, A. 1997. Starch Phosphorylase: Biochemical, Molecular and Biotechnological Aspects. *Biotechnol. Genet. Eng. Rev*. 17: 4.

Okogbenin, E., Setter, T. L., Ferguson, M., Mutegi, R., Ceballos, H., Olasanmi, B., and Fregene, M. 2013. Phenotypic approaches to drought in cassava: Review. *Front. Physiol*. 4: 1-15.

Oliver, S. N., Tiessen, A., Fernie, A. R., and Geigenberger, P. 2008. Decreased expression of plastidial adenylate kinase in potato tubers results in an enhanced rate of respiration and a stimulation of starch synthesis that is attributable to post-translational redox-activation of ADP-glucose pyrophosphorylase. *J. Exp. Bot*. 59: 315-325.

Olsen, K. M. and Purugganan, M. D. 2002. Molecular evidence on the origin and evolution of glutinous rice. *Genet*. 162: 941-950.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. 1999. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci*. 96: 2896-2901.

Patil, B. L. and Fauquet, C. M. 2009. Cassava mosaic Gemini viruses: Actual knowledge and perspectives. *Mol. Plant Pathol*. 10: 685-701.

Patle, S. and Lal, B. 2008. Investigation of the potential of agro-industrial material as low cost substrate for ethanol production by using Candida tropicalis and Zymomonas mobilis. *Biomass Bioenerg*. 32: 569-602.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Biochem.* 96: 4285-4288.

Petreikov, M., Shen, S., Yeselson, Y., Levin, I., Bar, M., and Schaffer, A. A. 2006. Temporally extended gene expression of the ADP-Glc pyrophosphorylase large subunit (AgpL1) leads to increased enzyme activity in developing tomato fruit. *Planta.* 224: 1465-1479.

Poolman, M. G., Miguet, L., Sweetlove, L. J., and Fell, D. A. 2009. A genome-scale metabolic model of Arabidopsis and some of its properties. *Plant physiol.* 151: 1570-1581.

Prestridge, D. S. 1995. Predicting Pol II Promoter Sequences Using Transcription Factor Binding Sites. *J. Mol. Biol.* 249: 923-932.

Prochnik, S., Marri, P. D., Desany, B., Rabinowicz, P. D., Kodira, C., Mohiuddin, M., Rodriguez, F., Fauquet, C., Tohme, J., Harkins, T., Rokhsar, D. S., and Rounsley, S. 2012. The cassava genome: current progress, future directions. *Trop. Plant Biol.* 5: 88-94.

Qiu, C., Yang, J., Ge, S., Chang, R., Xiong, L., and Sun, Q. 2016. Preparation and characterization of size-controlled starch nanoparticles based on short linear chains from debranched waxy corn starch. *LWT - Food Sci. Technol.* 74: 303-310.

Radchuk, V. V, Borisjuk, L., Sreenivasulu, N., Merx, K., Mock, H. P., Rolletschek, H., Wobus, U., and Weschke, W. 2009. Spatiotemporal profiling of starch biosynthesis and degradation in the developing barley grain. *Plant physiol.* 150: 190-204.

Regina, A., Berbezy, P., Kosar-Hashemi, B., Li, S., Cmiel, M., Larroque, O., Bird, A. R., Swain, S. M., Cavanagh, C., Jobling, S. A., Li, Z., and Morell, M. 2015. A genetic strategy generating wheat with very high amylose content. *Plant Biotechnol. J.* 13: 1276-1286.

Regina, A., Kosar-Hashemi, B., Ling, S., Li, Z., Rahman, S., and Morell, M. 2010. Control of starch branching in barley defined through differential RNAi suppression of starch branching enzyme IIa and IIb. *J. Exp. Bot.* 61: 1469-1482.

Rhee, S. Y., Beavis, W., Dixon, D., Huala, E., Weems, D. C., and Zhang, P. 2003. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.* 31(1): 224.

Rocca-Serra, P., Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Contrino, S., Vilo, J., Abeygunawardena, N., Mukherjee, G., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., and Sansone, S. 2003. ArrayExpress: a public database of gene expression data at EBI. *Comptes Rendus Biologies.* 326(10): 1075-1078.

Rolletschek, H., Hajirezaei, M. R., Wobus, U., and Weber, H. 2002. Antisense-inhibition of ADP-glucose pyrophosphorylase in Vicia narbonensis seeds increases soluble sugars and leads to higher water and nitrogen uptake. *Planta.* 214: 954-964.

Rongsirikul, O., Saithong, T., Kalapanulak, S., Meechai, A., Cheevadhanarak, S., Netrphan, S., and Suksangpanomrung, M. 2010. Reconstruction of starch biosynthesis pathway in cassava using comparative genomic approach. Proceedings of the first international conference on computational

systems-biology and bioinformatics, Bangkok, Thailand. Springer, Germany, pp. 118-129.

Rosti, S., Fahy, B., and Denyer, K. 2007. A mutant of rice lacking the leaf large subunit of ADP-glucose pyrophosphorylase has drastically reduced leaf starch content but grows normally. *Funct. Plant Biol.* 34: 480-489.

Saithong, T., Rongsirikul, O., Kalapanulak, S., Chiewchankaset, P., Siriwat, W., and Netrphan, S. 2013. Starch biosynthesis in cassava: a genome-based pathway reconstruction and its exploitation in data integration. *BMC Syst. Biol.* 7: 75.

Sakurai, T., Plata, G., Rodriguez-Zapata, F., Seki, M., Salcedo, A., and Toyoda, A. 2007. Sequencing analysis of 20,000 full-length cDNA clones from cassava reveals lineage specific expansions in gene families related to stress response. *BMC Plant Biol.* 7: 66.

Saripalli, G. and Gupta, P. K. 2015. AGPase: its role in crop productivity with emphasis on heat tolerance in cereals. *Theor. Appl. Genet.* 128: 1893-1916.

Satoh, H., Shibahara, K., Tokunaga, T., Nishi, A., Tasaki, M., Hwang, S. K., Okita, T. W., Kaneko, N., Fujita, N., Yoshida, M., Hosaka, Y., Sato, A., Utsumi, Y., Ohdan, T., and Nakamura, Y. 2008. Mutation of the Plastidial α-Glucan Phosphorylase Gene in Rice Affects the Synthesis and Structure of Starch in the Endosperm. *Plant Cell.* 20: 1833-1849.

Sawada, T., Nakamura, Y., Ohdan, T., Saitoh, A., Francisco, P. B., Suzuki, E., Fujita, N., Shimonaga, T., Fujiwara, S., Tsuzuki, M., Colleoni, C., and Ball, S. 2014. Diversity of reaction characteristics of glucan branching enzymes and the fine structure of α-glucan from various sources. *Arch. Biochem. Biophys.* 562: 9-21.

Seung, D., Soyk, S., Coiro, M., Maier, B. A., Eicke, S., and Zeeman, S. C. 2015. Protein targeting to starch is required for localising granule-bound starch synthase to starch granules and for normal amylose synthesis in Arabidopsis. *PLoS Biol.* 13: 1-29.

Shannon, J. C., Pien, F. M., and Liu, K. C. 1996. Nucleotides and Nucleotide Sugars in Developing Maize Endosperms (Synthesis of ADP-Glucose in brittle-1). *Plant physiol.* 110: 835-843.

Shobhit, J. A., Stamatoyannopolous, Timothy, B., and William, S. N. 2007. Quantifying similarity between motifs. *Genome Biol.* 8(2): 24.

Skeffington, A. W., Graf, A., Duxbury, Z., Gruissem, W., and Smith, A. M. 2014. Glucan water Dikinase exerts little control over starch degradation in Arabidopsis leaves at night. *Plant Physiol.* 165(2): 866-879.

Slewinski, T. L., Ma, Y., Baker, R. F., Huang, M., Meeley, R., and Braun, D. M. 2008. Determining the role of Tie-dyed1 in starch metabolism: Epistasis analysis with a maize ADP-glucose pyrophosphorylase mutant lacking leaf starch. *J. Hered.* 99: 661-666.

Smidansky, E. D., Meyer, F. D., Blakeslee, B., Weglarz, T. E., Greene, T.W., and Giroux, M. J. 2007. Expression of a modified ADP-glucose pyrophosphorylase large subunit in wheat seeds stimulates photosynthesis and carbon metabolism. *Planta.* 225: 965-976.

Smith, A. M. 2001. The biosynthesis of starch granules. *Biomacromolecules.* 2: 335-341.

Sonnewald, U. and Kossmann, J. 2013. Starches-from current models to genetic engineering. *Plant Biotechnol. J.* 11: 223-232.

Sreenivasulu, N., Radchuk, V., Strickert, M., Miersch, O., Weschke, W., and Wobus, U. 2006. Gene expression patterns reveal tissue-specific signalling networks controlling programmed cell death and ABA-regulated maturation in developing barley seeds. *Plant J.* 47: 310-327.

Srichuwong, S., Jane, J. L., William, M. 2007. Physicochemical properties of starch affected by molecular composition and structure: a review. *Food Sci. Biotechnol.* 16: 663-674.

Sumner, J. B., and Sommers, G. F. 1944. The water-soluble polysaccharides of sweet corn. *Arch. Biochem.* 4: 7-9.

Suzuki, E. and Suzuki, R. 2016. Distribution of glucan-branching enzymes among prokaryotes. *Cell. Mol. Life Sci.* 73: 2643-2660.

Taliercio, E. 2011. Characterization of an ADP-glucose pyrophosphorylase small subunit gene expressed in developing cotton (Gossypium hirsutum) fibers. *Mol. Biol. Rep.* 38: 2967-2973.

Tang, G. Q. and Sturm, A. 1999. The sucrose-cleaving enzymes of plants are crucial for development, growth and carbon partitioning. *Trends Plant Sci.* 4: 401-407.

Tang, X. J., Peng, C., Zhang, J., Cai, Y., You, X. M., Kong, F., Yan, H. G., Wang, G. X., Chen, X. G., Ma, J., Wang, P., Jiang, L., Zhang, W. W., and Wan, J. M. 2016. ADP-glucose pyrophosphorylase large subunit 2 is essential for storage substance accumulation and subunit interactions in rice endosperm. *Plant Sci.* 249: 70-83.

Taylor, N., Chavarriaga, P., Raemakers, K., Siritunga, D., and Zhang, P. 2004. Development and application of transgenic technologies in cassava. *Plant Mol. Biol.* 56(4): 671-688.

Tetlow, I. J. and Emes, M. J. 2014. A review of starch-branching enzymes and their role in amylopectin biosynthesis. *IUBMB Life.* 66: 546-558.

Tetlow, I. J., Bowsher, C.G., and Emes, M. J. 1996. Reconstitution of the hexose phosphate translocator from the envelope membranes of wheat endosperm amyloplasts. *Biochem. J.* 319(3): 717-723.

Thitisaksakul, M., Jimenez, R. C., Arias, M. C., and Beckles, D. M. 2012. Effects of environmental factors on cereal starch biosynthesis and composition. *J. Cereal Sci.* 56: 67-80.

Tiessen, A., Nerlich, A., Faix, B., Hümmer, C., Fox, S., Trafford, K., Weber, H., Weschke, W., and Geigenberger, P. 2012. Subcellular analysis of starch metabolism in developing barley seeds using a non-aqueous fractionation method. *J. Exp. Bot.* 63: 2071-2087.

Tiessen, A., Prescha, K., Branscheid, A., Palacios, N., McKibbin, R., Halford, N. G., and Geigenberger, P. 2003. Evidence that SNF1-related kinase and hexokinase are involved in separate sugar-signalling pathways modulating post-translational redox activation of ADP-glucose pyrophosphorylase in potato tubers. *Plant J.* 35: 490-500.

Timothy L. B., Nadya, W., Chris, M., Wilfred, W. L. 2009. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34(2): 369-373.

Tjaden, J., Möhlmann, T., Kampfenkel, K., Henrichs, G., and Neuhaus, H. E. 1998. Altered plastidic ATP/ADP-transporter activity influences potato (Solanum tuberosum L.) tuber morphology, yield and composition of tuber starch. *Plant J.* 16: 531-540.

Tsubone, W., Kubota, F., and Saltou, K. 1997. Effect of grafting on the activity of adenosine 5 diphosphate glucose pyrophosphorylase and tuberous root

production in sweet potato (Ipomoea batatas Lam). *J. Crop Sci.* 66: 509-510.

Tuncel, A. and Okita, T. W. 2013. Improving starch yield in cereals by over-expression of ADPglucose pyrophosphorylase: Expectations and unanticipated outcomes. *Plant Sci.* 211: 52-60.

Uchechukwu-Agua, A. D., Caleb, O. J., Manley, M., and Opara, U. L. 2015.Effects of storage conditions and duration on physicochemical and microbial quality of the flour of two cassava cultivars (TME 419 and UMUCASS 36). *J. Food.* 13: 1-11.

Utsumi, Y., Utsumi, C., Sawada, T., Fujita, N., and Nakamura, Y. 2011. Functional diversity of isoamylase oligomers: the ISA1 homo-oligomer is essential for amylopectin biosynthesis in rice endosperm. *Plant Physiol.* 156: 61-77.

Vanderschuren, H., Nyaboga, E., Poon, J. S., Baerenfaller, K., Grossmann, J., Hirsch-Hoffmann, M., Kirchgessner, N., Nanni, P., and Gruissem, W. 2014. Large-Scale Proteomics of the Cassava Storage Root and Identification of a Target Gene to Reduce Postharvest Deterioration. *Plant Cell.* 26: 1913-1924.

Verdier, V., Boher, B., Maraite, H., and Geiger, J. P. 1994. Pathological and molecular characterization of Xanthomonas campestris strains causing diseases of cassava (Manihot esculenta). *Appl. Environ. Microbiol.* 60: 4478-4486.

Vrinten, P.L. and Nakamura, T. 2000. Wheat granule-bound starch synthase I and II are encoded by separate genes that are expressed in different tissues. *Plant physiol.* 122: 255-264.

Wattebled, F., Dong, Y., Dumez, S., Delvallé, D., Planchot, V., Berbezy, P., Vyas, D., Colonna, P., Chatterjee, M., Ball, S., and D'Hulst, C. 2005. Mutants of Arabidopsis lacking a chloroplastic isoamylase accumulate phytoglycogen and an abnormal form of amylopectin. *Plant physiol.* 138: 184-195.

Wiberley-Bradford, A. E., Busse, J. S., Jiang, J., and Bethke, P. C. 2014. Sugar metabolism, chip color, invertase activity, and gene expression during long-term cold storage of potato (Solanum tuberosum) tubers from wild-type and vacuolar invertase silencing lines of Katahdin. *.BMC res. Notes.* 7: 801.

Xavier, M., Ruth, E., DomènecFarré, Oscar, N., and Javier, M. 2002. PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinforma.* 18(2): 333-334.

Xia, H., Yandeau-Nelson, M., Thompson, D. B., and Guiltinan, M. J. 2011. Deficiency of maize starch-branching enzyme I results in altered starch fine structure, decreased digestibility and reduced coleoptile growth during germination. *BMC plant biol.* 11: 95.

Yang, J., An, D., and Zhang, P. 2011. Expression profiling of Cassava storage roots reveals an active process of glycolysis/gluconeogenesis. *J. Integr. Plant Biol.* 53(3): 193-211.

Yin, Y. G., Kobayashi, Y., Sanuki, A., Kondo, S., Fukuda, N., Ezura, H., Sugaya, S., and Matsukura, C. 2010. Salinity induces carbohydrate accumulation and sugar-regulated starch biosynthetic genes in tomato (Solanum lycopersicum L. cv. 'Micro-Tom') fruits in an ABA-and osmotic stress-independent manner. *J. Exp. Bot.* 61: 563-574.

Yuang, Q., Ouyang, S., Liu, J., Suh, B., Cheung, F., Sultana, R., Lee, D., Quackenbush, J., and Buell, C. R. 2003. The TIGR Rice Genome

Annotation Resource: Annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res*. 31: 229-233.

Zeeman, S.C., Kossmann, J., and Smith, A. M. 2010. Starch: its metabolism, evolution, and biotechnological modification in plants. *Annu. Rev. Plant Biol.* 61: 209-234.

Zeng, F., Zhu, S., Chen, F., Gao, Q., and Yu, S. 2016. Effect of different drying methods on the structure and digestibility of short chain amylose crystals. *Food Hydrocoll.* 52: 721-731.

Zhang, L., Häusler, R. E., Greiten, C., Hajirezaei, M. R., Haferkamp, I., Neuhaus, H.E., Flügge, U. I., and Ludewig, F. 2008. Overriding the co-limiting import of carbon and energy into tuber amyloplasts increase the starch content and yield of transgenic potato plants. *Plant Biotechnol. J.* 6: 453-464.

Zhou, Y. X., Chen, Y. X., Tao, X., Cheng, X. J., and Wang, H. Y. 2016. Isolation and characterization of cDNAs and genomic DNAs encoding ADP-glucose pyrophosphorylase large and small subunits from sweet potato. *Mol. Genet. Genomics*. 291: 609-620.

APPENDICES

8. APPENDIX I

8.1 E2P2 (Ensemble Enzyme Prediction Pipeline Version 3.0) SOFTWARE INSTALLATION

The archived package is available from: https://dpb.carnegiescience.edu/labs/rhee-lab/software

Unzip and extract the E2P2 package in your target location:

tar -xzf e2p2-3.0.tar.gz

8.2 E2P2 (Ensemble Enzyme Prediction Pipeline Version 3.0) SOFTWARE OPERATION

E2P2 was built to run on 64-bit Linux systems.

The input file containing protein sequence data must be in FASTA format.

Specify the paths of the input and output files and run E2P2 with the following command:

./runE2P2.v3.1.py -i <input filename> -o <output filename>

'pf-EC-to-official-RXN.pl' is already built-in and needed as the final step of runE2P2.py.

"COMPARATIVE AND FUNCTIONAL GENOMICS ANALYSIS OF STARCH BIOSYNTHESIS PATHWAYS IN CASSAVA"

By

POOJA HARSHAN

(2012-09-102)

Abstract of Thesis

Submitted in partial fulfilment of the

Requirement for the degree of

MASTER OF SCIENCE (INTEGRATED) IN BIOTECHNOLOGY

Faculty of Agriculture
Kerala Agricultural University, Thrissur

B. Sc. - M. Sc. (INTEGRATED) BIOTECHNOLOGY

DEPARTMENT OF PLANT BIOTECHNOLOGY

COLLEGE OF AGRICULTURE

VELLAYANI, THIRUVANANTHAPURAM - 695 522

KERALA, INDIA

2017

9. ABSTRACT

The study entitled "Comparative and Functional Genomics Analysis of Starch biosynthesis Pathways in Cassava [*Manihot esculenta* Crantz.]" was carried out at the Section of Extension and Social Sciences, ICAR-Central Tuber Crops Research Institute, Sreekariyam, Thiruvananthapuram during 2016-2017. The objectives of the study were to study the starch biosynthesis pathways in cassava using data integration and also the temporal gene expression pattern during different growth stages. The knowledge on differentially expressed genes can be exploited for plant breeding programs in order to develop high starch cassava varieties.

Starch related gene and protein sequences were collected from the respective databases of cassava and each of the template plants *viz.* Arabidopsis, Maize, Rice, Castor bean and Potato. Reciprocal Blast was performed to check for respective orthologues that corresponds to the template plants in cassava with a predetermined function. Three novel gene sequences were predicted electronically with functions similar to alpha 1, 4 glucan branching enzyme, sucrose phosphate synthase and UDP glycosyl-transferase super family. The identified gene sequences were annotated using E2P2 (Ensemble Enzyme Prediction Pipeline) software which is maintained by the Plant Metabolic Network (PMN). Protein Motif Analysis (PMA) was carried out using the MEME Suite version 4.12.0 which is a package of tools for motif analyses like motif discovery, motif enrichment, motif scanning and motif comparison.

Starch biosynthesis pathways in cassava include the carbon-dioxide fixation, starch and sucrose biosynthesis pathways as a whole. So in order to get a better comprehension on various pathways that regulate starch metabolism in the crop, the entire starch biosynthesis pathways were made into a single consolidated pathway. This reduces the complexity of pathway reconstruction for different pathways yielding the same end product. Pathway reconstruction becomes

beneficial to the agricultural community only when some useful information on the genes that are linked to some phenotypic traits is considered in the pathway. Insights on the cis regulatory elements involved in the pathway enhance the possibility of starch production and accumulation in the roots of cassava. Promoter analysis and Transcription factor prediction was carried out in the work. The tool employed for promoter analysis is Promoter Scan maintained at BIMAS (Bioinformatics and Molecular Analysis Section), NIH. For transcription factor and transcription factor binding site prediction, PlantTFDB (Plant Transcription Factor Database) was used.

Candidate gene prioritization is one among the other approaches being used in the study. Through gene prioritisation, five potential candidate genes have been found out which have a major role in revolutionising the starch biosynthesis pathway in cassava. One of the genes was a negative regulator/ suppressor of starch synthesis in cassava which made a breakthrough in the work. Finally, the output data of all the approaches were integrated together in the pathway constructed to create a better interactive pathway for visualization. Cell illustrator version 4.0 was employed for the construction of the starch biosynthesis pathway and output data integration in cassava. Hence, the starch biosynthesis pathway was not merely reconstructed but also the molecular insights into the genes and other auxiliary regulatory elements were also incorporated to create a highly interactive pathway.

174176