"*IN SILICO* ANALYSIS OF CAROTENOID BIOSYNTHESIS PATHWAY IN CASSAVA (*Manihot esculenta* C.)"

By

MUHAMMED SADIQ P. A.

(2012-09-119)

Thesis

Submitted in partial fulfilment of the

Requirement for the degree of

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

Faculty of Agriculture
Kerala Agricultural University, Thrissur

DEPARTMENT OF PLANT BIOTECHNOLOGY

COLLEGE OF AGRICULTURE

VELLAYANI, THIRUVANANTHAPURAM - 695 522

KERALA, INDIA

2017

DECLARATION

I hereby declare that this thesis entitled "***In silico*** **analysis of carotenoid biosynthesis pathway in cassava (*Manihot esculenta* C.)**"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.

Vellayani

Date: 03/11/2017

MUHAMMED SADIQ P. A.

(2012-09-119)

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 ***"In silico* analysis of carotenoid biosynthesis pathway in cassava (*Manihot esculenta C.)*"** is a record of research work done by Mr. Muhammed Sadiq P. A. (2012-09-119) under my guidance and supervision and hat this is not previously formed the basis for the award of any degree, diploma, fellowship or associateship to him.

Sreekariyam
Date: 31.10.2017

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

Phone : 91-0471-2598551 to 2598554
Director (Per.) : 91-0471-2598431
(Res.) : 91-0471 - 2441957
Admn.Officer : 91-0471-2598193

Fax : 91-0471-2590063
E-Mail : ctcritvm@yahoo.com
ctcritvm@gmail.com
Web : http://www.ctcri.org

CERTIFICATE

We, the undersigned members of the advisory committee of Mr. Muhammed Sadiq P. A. a candidate for the degree of B. Sc.- M. Sc. (Integrated) Biotechnology, agree that the thesis entitled "***In silico* analysis of carotenoid biosynthesis pathway in cassava (*Manihot esculenta* C.)** may be submitted by Mr. Muhammed Sadiq P. A. in partial fulfilment of the requirement for the degree.

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

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

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

Dr. Saravanan Raju
(Member, Advisory Committee)
Senior scientist
Division of Crop Utilization
ICAR-CTCRI, Sreekariyam, Trivandrum

EXTERNAL EXAMINER

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

ACKNOWLEDGEMENT

It is with my heartfelt feelings, I wish to express my deep sense of gratitude and sincere thanks to Dr. J. Sreekumar, Principal Scientist, Section of Extension and Social Sciences, ICAR-CTCRI, Sreekariyam, Thiruvananthapuram and his valuable guidance, sincere help, patience, support and encouragement throughout the course of study.

I am thankful to the members of the advisory committee, Dr. Swapna Alex, Professor and Head, Department of Plant Biotechnology, College of Agriculture, Vellayani, Dr. C Mohan, Principal Scientist, Division of Crop Improvement, ICAR-CTCRI, Sreekariyam, Thiruvananthapuram, Dr. Saravanan Raju, Senior Scientist, Division of Crop Utilization, ICAR-CTCRI, Sreekariyam, Thiruvananthapuram.

I am indebted to all my teachers for giving me the confidence to walk ahead. I am thankful to Dr. K. B Soni, Dr. R. Beena and Dr. A. N Jyothi for their critical professional advice and direction throughout my research.

My whole hearted thanks to all my dear friends and seniors- Adarsh Prathap, Vipin Mohan, Ajin Sekhar, Pooja Harshan, Ambu Vijayan, Selvi Sree, for guiding me and making my study memorable.

At last I owe this achievement to my parents, who always stood along my side and I will never forget the timely help, mental support, kindness and affection extended by other

family members without them this work would have never seen light.

I acknowledge the favour of numerous persons who, though not been individually mentioned here, who have all directly or indirectly contributed to this thesis work.

MUHAMMED SADIQ P. A.

TABLE OF CONTENTS

Sl. No.	Chapters	Page No.
1	INTRODUCTION	1
2	REVIEW OF LITERATURE	3
3	MATERIALS AND METHODS	24
4	RESULTS	39
5	DISCUSSION	53
6	SUMMARY	59
7	REFERENCES	61
8	ABSTRACT	72

LIST OF TABLES

Table No.	Title	Page no.
1	HfPNet elements in cell illustrator	37
2	Genes involved in carotenoid biosynthesis pathway in cassava	41
3	E2P2 functional annotation of carotenoid genes	43
4	Transcription factors involved in carotenoid pathway in cassava	44
5	CpG islands in carotenoid genes	46
6	Unusual Domains of carotenoid proteins in Cassava	48
7	miRNA targeting carotenoid genes in cassava	49
8	The expression profile for cassava carotenoid genes	51

LIST OF FIGURES

Figure No.	Title	Btw Page
1	Workflow for the *in silico* analysis of carotenoid biosynthesis pathway in Cassava	25-26
2	Number of transcription factors regulating carotenoid genes in cassava	44-45
3	Number of transcription factors related to stress in cassava	44-45
4	Number of transcription factors related to phytohormones in cassava	44-45
5	Number of transcription factors related to growth and development of cassava	44-45
6	Overall functions of transcription factors regulating carotenoid genes in cassava	44-45
7	Conserved domains of carotenoid protein sequences in cassava	47-48
8	Constructed phylogenetic tree and gene duplications of carotenoid genes in cassava using MEGA	49-50
9	Constructed phylogenetic tree and gene duplications of carotenoids in template plants and cassava using MEGA	49-50
10	Chromosomal location analysis of carotenoid genes in cassava using NCBI Map viewer	49-50
11	Number of expressed carotenoid genes in cassava retrieved from NCBI EST database	50-51
12	Screen shot- Subcellular localization of carotenoid genes in cassava using Targetp	51-52
13	Constructed putative carotenoid biosynthesis pathway in cassava using Cell illustrator	60-61

LIST OF ABBREVIATIONS AND SYMBOLS USED

ABA	Abscisic Acid
AMF	Arbuscular Mycorrhizal Fungi
AR	Adventitious Root
ATIPD	*Arabidopsis thaliana* Isoprenoid Database
BioPax	Biological Pathway Exchange
BLAST	Basic Local Alignment Search Tool
CCD	Carotenoid Cleavage dioxygenase
CDD	Conserved Domain Database
CDART	Conserved Domain Architecture Retrieval Tool
CI	Cell Illustrator
CRITSO	Carotenoid Isomerase
CSML	Cell System Mark-up Language
DBD	DNA Binding Domain
DDBJ	DNA Data Bank of Japan
dbEST	Database for Expressed Sequence Tag
E2P2	Ensemble Enzyme Prediction Pipeline
ENA	European nucleotide Archive
EST	Expressed Sequence Tag
GEO	Gene Expression Omnibus
KEGG	Kyoto Encyclopaedia of Gene and Genomes
lncRNA	Long Non-Coding RNA

LR	Lateral Root
LYC	Lycopene Cyclase
MEGA	Molecular Evolutionary Genetic Analysis
MEP	Methyl Erythritol Phosphate
miRNA	microRNA
MMDB	Molecular Modelling Database
NCBI	National centre For Biotechnology Information
NCED	Epoxycarotenoid Dioxygenase
PATIKA	Pathway Analysis Tools For Integration and Knowledge Base
PDS	Phytoene Desaturase
PGDB	Pathway Genome Database
PGSC	Potato Genomic Sequence Consortium
PlantRegMap	Plant Regulatory Map
PMN	Plant Metabolic Network
PMC	PubMed Central
PRIAM	PRofils pour l'Identification Automatique du Métabolisme
PSY	Phytoene Synthase
PSSM	Position Specific Scoring Matrix
PSI-MI	Proteomics Standard Initiative Molecular Interaction
psRNATarget	A Plant Small RNA Target Analysis Server
RPSD	Reference Protein Sequence Dataset
RPS-BLAST	Reversed Position Specific Blast

SAVI	Semi-Automated Validation and Integration Pipeline
SBML	System Biology Mark-up Language
SL	Strigolactone
TAIR	The Arabidopsis Information Resource
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TFDB	Transcription Factor Database
TR	Transcriptional Regulators
XML	eXtensible Markup Language
ZDS	Zeta Carotene Desaturase
ZISO	Zeta Carotene Isomerase

INTRODUCTION

1. INTRODUCTION

Cassava (*Manihot esculenta* Crantz) has its origin in Latin America where the indigenous Indian population has cultivated it for at least 4000 years. Cassava, also called manioc or tapioca is a perennial woody shrub which is cultivated for its starchy storage roots throughout tropical and subtropical regions of the world particularly in South America, Africa, and Asia. It is the key provenance of dietary energy for more than 500 million people. Cassava production in Asia was augmented at a high rate of 3% annually during the late 70s and early 80s, slowed down during the 90s, and has been mounting quite rapidly again at 5.6% per year during the past ten years, and at a very high rate of 9.1% during the past 5 years. Africa now produces more cassava than the rest of the world and it is one among the continent's staple food crops.

Carotenoids are the second most naturally arising pigments on earth with additional 750 members. They are C_{40} lipophilic isoprenoids produced in all photosynthetic organisms. Carotenoids and its oxidative and enzymatic cleavage products called apocarotenoids are vital for numerous biological progressions in plants such as signalling pathways, photosynthesis and photo protection and assist in growth and development. In cassava, the total carotenoid and β-carotene content ranges between 2.64-14.15 µg/g and 1.99-10.32 µg/g seen predominantly in roots.

Many approaches are there for improving the quality, nutritional content and cassava production. Use of bioinformatics and other high throughput technologies are beneficial for improving the quality and production in cassava through gene manipulation at the genome level. Approaches like comparative genomics, gene analysis and biosynthetic pathway analysis play a major role in the improvement of cassava. Comparative genomics approaches aim at solving complex biological problems in carotenoid biosynthesis will be useful to develop new biotechnological strategies in crop improvement programs. Acquiring the precise biology of agriculturally important traits governed by complex gene regulatory networks and pathways will enable us to develop a smart cassava crop. Smart crop is a portion of agricultural novelty that yields more value-added products in a smaller period. It

reduces our dependence on pesticides and fungicides thereby supporting eco-friendly as well as sustainable agriculture.

Progression in the era of bioinformatics from heuristic approaches to high through put plant genome sequencing revolutionized plant metabolic pathway studies. Development of recent pathway construction methodologies like Design-construction-evaluation-optimization (DCEO) biotechnology will be useful for design, construction, evaluation, and optimization of different biological pathways at the systems level (Chen *et al.*, 2017). The tight association of genes between metabolic pathways in plants is revealed by global co-expression network approach.

The development of molecular regulation, interaction maps and incorporation of regulatory elements in multi-layered networks reduces the gene network complexity (Wong *et al.*, 2017). All these methods offer a theoretical and technological outline to exploit, alter and create new pathways in plants and other organisms. Hence, the present study aimed at identifying the key genes and the transcriptional regulatory elements participating in the carotenoid metabolism and thereby constructing a consolidated pathway of carotenoid biosynthesis in cassava.

REVIEW OF LITERATURE

2. REVIEW OF LITERATURE

2.1 CAROTENOIDS

Carotenoids are one of the substantial categories of plant pigments and play a dynamic part in determining the quality constraints of fruit and vegetables. They are seen predominantly in algae, plants and photosynthetic bacteria. Some 600 diverse carotenoids are recognized to ensue naturally and new carotenoids continue to be recognized. They are responsible for the coloration after the leaf chlorophyll has been demolished. There are two most important classes of carotenoids i.e., carotenes and xanthophylls. Carotenes are hydrocarbons that can be cyclized at one or both the final ends of the molecule and xanthophyll is the oxygenated derivative of carotene. There is also an alternative group of carotenoids called apocarotenoids which are the cleavage products of carotenoids (Cazzonelli *et al.*, 2010).

Lutein (30-60%) is the utmost rich carotenoid in plant followed by β-carotene with (25-30%), neoxanthin and violaxanthin (10-15%). The quantity of carotenoids in leucoplast is very low that it is indistinguishable to human eye (Ruiz-Sola *et al.*, 2012). Chromoplasts snatch huge dimensions of carotenoids in plastoglobules and carotenoid storage assemblies of profiles composed of lipids and proteins (Deruere *et al.*, 1994).

The isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) were the precursors of carotenoids. Two self-controlling pathways occur in plant cells for the synthesis of these two prenyl diphosphate predecessors but the carotenoids are chiefly shaped from IPP and DMAPP molded by the plastidial 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway (Rodriguez-Concepcion, *et al.*, 2010). Accumulation of three IPP molecules to DMAPP creates geranyl-geranyl diphosphate (GGPP), a pioneer of numerous sets of plastidial isoprenoids as well as carotenoids (Lichtenthaler, *et al.*, 2001). The primary dedicated phase of carotenoid biosynthesis pathway is the making of 40-carbon phytoene from the condensation of two molecules of GGPP. This reaction is catalysed by the enzyme phytoene synthase

(PSY), which is the main bottleneck in the carotenoid pathway (Hirschberg *et al.*, 2001).

Phytoene are desaturated to produce the chromophore contains conjugated double bond chains that produce the plant carotenoids pillar and fixes their corporal and biotic assets. Desaturation and isomerization of uncoloured phytoene ultimately leads to the production of lycopene, a red carotenoid. The cyclization of the ends of lycopene-polyene chain is the primary branching point in the pathway resulting in the production of carotenes either with one ß ring and one ε ring (α-carotene) or with two ß rings (ß-carotene). Carotenoids with two ε rings do not occur in Arabidopsis and are unusual in plants. Hydroxylation of the carotene rings produces xanthophylls such as lutein (from α-carotene) and zeaxanthin (from ß-carotene). Zeaxanthin is epoxidated twice to make violaxanthin which can be later modified to produce neoxanthin. These carotenes and xanthophylls can be more altered to create an extensive variety of carotenoids constituents in plants and other organisms (Walter and Strack 2011).

In addition, carotenoids can be cleaved to create apocarotenoids, a structurally different class of compounds extensively distributed in nature. Other apocarotenoids are tangled in association of plants with their biological environment serving as pigments and flavours that invite pollinators or seed-distributing animals (Giuliano *et al.*, 2003).

Carvalho *et al.* (2016) reported that carotenoid accumulation in cassava has been linked with color of storage roots. They classified carotenoid sketches and ample of vital transcripts allied with biosynthesis of carotenoids. Among 23 cassava landraces, storage root extending in colour between white to yellow to pink, only five carotenoid categories were discovered in root with white colour and between 1 to 21 in storage root showed pink and yellow colours. The mainstream of storage root in these landraces stretched in colour between pale to penetrating yellow. Among this colour group, the total beta carotene enclosing all-E-, 9-Z and 13-Z-β-carotene isomers was the chief carotenoid category sensed fluctuating from 26.13% to 76.72 %.

Even though no α-carotene was detected, adjustable amounts of a α-ring resultant xanthophyll, lutein was identified with superior accumulation of α-ring xanthophylls than of β-ring xanthophyll. Lycopene was noticed in a landrace (Cas51) with pink color storage root but was not identified in storage root with yellow color.

2.2 GENES AND ENZYMES IN CAROTENOID PATHWAY

2.2.1 MEP pathway

The MEP pathway also represented as the 1- deoxy-D-xylulose 5-phosphate (DXP) pathway as well as precursor pathway of carotenoids produces IPP and dimethylallyl diphosphate (DMAPP) in plastids. The MEP pathway fits with the condensation of pyruvate and D-glyceraldehyde 3-phosphate (GA-3P). The condensation reactions were catalysed by 1-deoxy-D-xylulose 5-phosphate synthase discharges CO_2 and is an irreversible reaction that obligates carbon to the MEP pathway. In most plant species, DXS is specified by multiple gene paralogs. All further MEP-pathway enzymes in many other plant species are prearranged by one-copy genes.

The intramolecular redisposition and decline of DXP to MEP was catalyzed by DXP reductoisomerase. MEP is then transformed to 4-(cytidine 5 -diphospho)-2-C-methyl-Derythritol (CDP-ME) in a CTP-reliant process catalysed by CDP-CT. The hydroxyl cluster in the C_2 spot of CDP-ME was more phosphorylated by 4-(cytidine 5 - diphospho)-2-C-methyl-D-erythritol kinase. The artefact 2-phospho-4-(cytidine 5 -diphospho)- 2-C-methyl-D-erythritol (CDP-ME2P) is then transformed to 2-C-methyl-Derythritol 2,4-cyclodiphosphate (MEcPP), which is catalysed by 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase(MDS). MEcPP is then shortened by 4-hydroxy-3-methylbut-2-enyldiphosphate (HMBPP) synthase (HDS) to HMBPP, which is to end altered by HMBPP reductase (HDR) into a mixture of IPP and DMAPP (Vranova *et al.*, 2013)

2.2.2 IPP isomerase

The primary dedicated phase of the pathway is the development of phytoene from two GGPP molecules. DMAPP was the primary triggered substrate for the development of elongated polyisoprenoid complexes such as GGPP. The

establishment of DMAPP from IPP is a rescindable process that is catalysed by the IPP isomerase (Ruiz-sola *et al*., 2012).

2.2.3 GGPP synthase

The 20-carbon geranyl geranyl pyro phosphtae, functioning as the instant forerunner for carotenoid biosynthesis is formed by consecutive and lined accumulation of three IPP molecules and one DMAPP molecule. The Arabidopsis genome comprises of a family of additional 10 genes encoding recognized GGDS isoforms (Lange and Ghassemian 2003).

2.2.4 Phytoene synthase

The primary devoted phase in the biosynthesis of plant carotenoids are the production of phytoene from GGPP. This reaction catalysed by PSY, which is possibly the best-examined enzyme of the plant carotenoid family. The PSY enzyme catalyses a two-step reaction, the face-to-face abridgment of two GGPP molcules to develop the reaction intermediary pre-phytoene diphosphate trailed by the abolition of the diphosphate group from this intermediate in a complex rearrangement that comprises a carbocation neutralization to create phytoene (Dogbo *et al*., 1988). Although only one PSY gene (At5g17230) exists in arabidopsis, minor gene families seem to encrypt PSY in furthermost of the plants. For illustration, two genes encode PSY in tobacco, three genes found in tomato, maize, rice and cassava (Fraser *et al*., 1999). Arango *et al*. (2010), discovered three PSY genes in cassava of which one (MePSY3) develop a distinct division with the stress-aligned Poaceae homologs. Nevertheless, MePSY3 copies were nearly lacking in most of the tissues examined and unaltered upon abiotic stress action. In difference, the two lasting PSY genes donated dissimilarly to biosynthesis of carotenoids in roots, leaves and flower organs and reacted to drought and salt-stress circumstances.

2.2.5 Desaturases

The four consecutive desaturation reactions from phytoene to lycopene are catalysed by multiple associated enzymes in plants namely phytoene desaturase (PDS) and zeta carotene desaturase (ZDS). The enzymes generate extra double bonds in the

C_{40} backbone of phytoene thereby altering the colourless carotenoid into a coloured compound. Tanaka *et al*. (2008) reported that the desaturation steps consecutively convert phytoene into phytofluene, ζ-carotene, neurosporene and lycopene cumulating the sum of conjugated double bonds to five, seven, nine and eleven respectively. The two desaturations of phytoene to produce zeta-carotene via phytofluene were catalysed by (PDS) and the two desaturations of zeta-carotene to yield lycopene via neurosporene that was catalysed by zeta-carotene desaturase (ZDS). Solitary genes encrypt for PDS and ZDS in Arabidopsis (At4g14210 and At3g04870 respectively) and most other plants (Dong *et al*., 2007).

2.2.6 Isomerases

The enzyme catalyses the cis-trans isomerisation of the 15-15 carbon-carbon double bond in 9, 15, 9'-tricis-zeta-carotenoid which is essential for biosynthesis of entire plant carotenoids. Z-ISO is a plastid-targeted enzyme anticipated to be a vigorous membrane protein with many membrane-spanning domains (Zybailov *et al*., 2008; Ishikawa *et al*., 2009). The solitary Arabidopsis gene encrypting Z-ISO (At1g10830) yields multiple different transcripts (Chen *et al*., 2010). Later the assembly of 9,9'-di-cis-ζ-carotene by Z-ISO, the enzyme ZDS changes it into 7,9,7',9'-tetra-cis-lycopene (pro-lycopene) via 7,9,9'-tri-cis-neurosporene. CRTISO catalyses the cis–trans reactions to isomerise the four cis-bonds and develops as a chief supervisory nodule in the pathway. The determination of these cis-carotenoids is mainly unidentified and negligent to yield a chance that may display a portion as new signalling molecule. In recent times, CRTISO exposed to be vital for launching asymmetry among cis and trans carotenoid isomers (Yu *et al*., 2011). CRTISO enzymes display less sequence comparisons to plant desaturases (PDS, ZDS) and, to a greater degree to the bacterial *crtI* enzymes (Sandmann *et al*., 2009).

2.2.7 The cyclases

Lycopene beta-cyclase catalyses the growth of the bicyclic beta-carotene as of lycopene in plants and cyanobacteria. This enzyme grants two β -rings at the tip of the linear lycopene molecule. In the photosynthetic set of plants carotenoids were seen as bicyclic compounds, most recurrently with two β or altered βring. In plants and

cyanobacteria. the lycopene β cyclase (LCYB) catalyses the creation of the bicyclic β-carotene commencing the direct proportioned lycopene Neurosporene was brought as the substrate. Udoh *et al.* (2017), found that genetic enhancement of carotenoid amount in cassava conceivable through detecting single nucleotide polymorphism (SNP) accredited to disparity in carotenoid concentration between some cassava genotypes. Nevertheless, in contrast to the β-cyclase, the ε-cyclase incorporate only single ring to the proportioned lycopene, creating the monocyclic δ-carotene. When pooled, the β and ε-cyclases transform lycopene to form the alpha-carotene (β, ε-carotene), a carotenoid with single β and one ε ring that assists as the forerunner for development of lutein.

Song, *et al.* (2016), reported that protein sequence of potato LCYB presented significant resemblance LCYbs in many species and was articulated in all tissues and the peak amount was perceived in tubers and the lowest level was in roots. This also shows about 1.9 times increment in beta carotene content of transgenic potato tubers. This results also gives an insight to the beta carotene enrichment process of cassava.

2.2.8 The hydroxylases

The thylakoid membranes of plants encompass maximum of the carotenoid pigments, especially xanthophylls. Hydroxylation at the third carbon in each ring of the hydrocarbons β-carotene and α-carotene will yield the renowned xanthophyll pigments such as zeaxanthin and lutein correspondingly. Genetic indication and practical examination of an Arabidopsis beta-hydroxylase nourishes the presence of discrete hydroxylases exactly for the beta and zigma rings. Kang *et al.* (2017), reported that suppression of β-carotene hydroxylase gene increased the β-carotene content and tolerance to abiotic stress in transgenic sweet potato plants.

2.2.9 β-C-4-oxygenase

Carotenoid oxygenases are the group of enzymes intricated in the cleavage of carotenoids to yield retinal usually known as vitamin A. Carotenoids using keto classes in the four positions of the rings are extensively disseminated. Accumulation of a keto cluster at the fourth place of individual or both rings of the yellow β-carotene will yield the reddish-orange to red pigments echinenone and canthaxanthin.

2.2.10 Epoxidase and de-epoxidase

Zeaxanthin epoxidation into violaxanthin *via* antheraxanthin and violaxanthin de-epoxidation to reinforce zeaxanthin mentioned as xanthophyll, epoxide cycle 0r violaxanthin. The ZEP of pepper was institute to necessitate ferredoxin and ferredoxin: NADP oxidoreductase in count to NADPH, thus connecting condensed ferredoxin as the reductant for the epoxidase reaction. Jorge *et al.* (2016), reported that zeaxanthin epoxidase (ZEP) was the chief provider for carotenoid content by means of mutants missing ZEP action displaying a notable 6-fold upsurge in entire seed carotenoids comparative to the wild type. Natural disparity in ZEP gene expression in the course of seed progression was recognized as the fundamental appliance for the refinement of carotenoid composition, stability, and eventually content in arabidopsis seed.

2.2.11 Epoxycarotenoid cleavage enzyme

The epoxy carotenoid neoxanthin and violaxanthin have been believed to be forerunners for biosynthesis of the ABA.

2.3 APOCAROTENOIDS

The cleavage products of carotenoids are called apocarotenoids. These was produced by an oxidative cleavage of carotenoids by double bond specific cleavage enzyme (CCD). Cazzonelli *et al.* (2010), reported that this apocarotenoids are compounds rich in diverse functions like they act as colorants, antioxidants, hormones, signalling components, chromophores and scent/aroma constituents. Apocarotenoids are the precursors of plant hormones like ABA and strigolactones. Carotenoids also play an important role in photosynthesis by reducing photo oxidative damage and transferring absorbed lights to the photo system. The protection was modulated by the chloroplast containing carotenoid pigments by stabilizing membranes and acting as accessory and physical pigments in photosystem. Nishino *et al.* (2002), found that it is also vital to anthropoid well-being by existing as the forerunners of retinol that act as antioxidants and play a vibrant part in dropping age related macular degeneration and reducing hazard of various cancers.

Farmer *et al.* (2013), reported that majority of the apocarotenoids comprise a β-unsaturated carbonyl molecule able to effortlessly respond with the nucleophilic moieties of biological molecules and in the similar background. Ramel *et al.* (2011), stated that light stress persuades the oxidation of the β-carotene, priming to the accumulation of different volatile derivatives. One such compound, β-cyclocitral, was found to persuade variations in the expression of a big set of genes that have been recognized as O_2 responsive genes. The outcomes designate that β-cyclocitral is a stress signal formed in high light that is talented to persuade defence mechanisms and characterizes a prospective messenger intricated in the O_2 signalling pathway in plants. These volatile derivatives or reactive species developing from apocarotenoids and carotenoids can act with thiols on TFs, varying expression of gene (levonen *et al.*, 2004). Fester *et al.* (2002a), testified that CCDs had a role in non-volatile production, production of mycorradicin in medicago. The yellow pigment accrues to maximum heights in the roots of several plant classes on colonization with arbuscular mycorrhiza (AM) through root carotenoids cleavage processing. Burns *et al*. (2012), testified that AMF colonization with cassava roots increased total nutritional quality of the plant and AMF should be well-thought-out in exertions looking to progress the nourishment of plants in emerging world agricultural systems.

The cellular localization of CCDs and their substrates are subject to the production of apocarotenoid. In this background, Floss *et al*. (2008), described that CCD1 enzymes cleaves those apocarotenoids shaped by other CCDs proficient of dispersing out of the plastid. Vogel *et al.* (2008) suggested that carotenoid escape from plastids happen through insect attack making insect repellent compounds. This concept harmonizes with the opinion that CCD1 over-expression can decrease crucifer beetle attack (Utama *et al.,* 2002).

Apocarotenoids show their diversity in their nature by degradation with different CCD enzymes because these apocarotenoid enzymes involved in majority of the plants attacks the precursor carotenoids at different double bonds. Thus, diversity in the cleavage of carotenoids result in different apocarotenoids with diverse functions like aroma, nutrition, color and resistance to abiotic and biotic stress. Apocarotenoid influences a wide variety of biological processes in plants. In general, CCD1 and

CCD7 codes for β-ionone, this having a function of pollinator attractant and flavouring component (Kiefer *et al.,* 2001). Caceres *et al.* (2016), described that over expression of CCD1 gene results herbivores to be deterred or engrossed to over expressing plants comparative to the wild type. In Arabidopsis, bioassays confirmed that β ionone has robust repellent outcome to both flea beetle, two spotted spider mites (*Tetranychus urticae*), and silverleaf whiteflies (*Bemisia tabaci*). These findings reveal that how modifiable genes of the carotenoid pathway can upsurge herbivore deterrent volatiles, a new tool for insect pest supervision. CCD2, CCD3, CCD4, CCD6 and CCD9 codes for abscisic acid, phytohormone required for dormancy and drought tolerance (Schwartz *et al*., 1997; Tan *et al*., 1997; Tan *et al*., 2003). CCD7 and CC8 codes for MAX product, plant growth regulator (Booker *et al*., 2004). Fester *et al.* (2005), reported that CCD1 codes for both blumenin and mycorradicin, having antifungal properties.CCD7 and CCD8 are intricated in the formation of strigolactones from β-carotene *via* carlactones.

Strigolactones (SLs) and their derivatives were lately well-defined as novel phytohormones that compose shoot and root growth. Stages of SLs, which are shaped mainly by plant roots, upsurge under low nitrogen and phosphate levels to regulate plant retorts (Sun *et al*., 2016). These apocarotenoids can play a vital role in root crop like cassava. SL is one of the apocarotenoids involved in plant growth and development. The genes involved in apocarotenoid biosynthesis are already identified in cassava but it is not well studied. Accumulating signs specifies that SLs take part in root growth in numerous plant species.

SLs encourage the elongation of primary roots and adventitious roots (ARs) and block lateral root (LR) formation (Kapulnik, *et al*., 2011). In accumulation to SLs, auxins show a crucial role in launching designs of root morphology and was regulated by the stages of N and P. As a benefit to our study, arabidopsis apocarotenoid data specifies that, there are nine CCDs characterized rendering to their enzymatic peculiarities and they govern the CCD nomenclature in other species. NCED 2, 3, 5,6 and 9 involved in ABA biosynthesis (Tan *et al*., 2003).CCD1,4, and 7 have extensive substrate particularity while CCD8 was involved in strigolactone biosynthesis. Single extra CCD subclass, named CCD2 also recognized, in Crocus sativus, it yields

25

saffron-flavour and scent (Alder *et al.*, 2012). Apocarotenoids, like β-cyclocitral, β-ionone, geranial, geranyl acetone, theaspirone, α-damascenone and β-damascenone, altogether subsidize to the flavour and/or fragrance of flowers and a diversity of foods. Their assemblies disclose an isoprenoid-based origin, and they were extensively expected products of the oxidative cleavage of carotenoids.

2.4 DATASETS AND TOOLS

2.4.1 Phytozome database

Phytozome (http://www.phytozome.net) was available in 2008. It is a central pivot that license handlers with changing grades of computational complexity to approach functionally assigned plant gene categories, to direct the evolutionary antiquity of gene folks and distinct genes, to scrutinize plant genes in their genomic background, to allocate assumed role to uncharacterized user sequences and delivers unvarying admission to plant genomics data sets comprising of whole genomes, gene and associated (e.g. homologous) sequences and alignments, gene functional info and gene families, either in unpackaged or as the outcome of on the fly multifaceted queries. The Phytozome web gateway assimilates a sum of extensively used open source apparatuses for gene family exploration, examination and estimation (Goodstein *et al*., 2012)

2.4.2 NCBI database

The National Centre for Biotechnology Information (NCBI) at the National Institutes of Health was established in 1988 to develop information systems for molecular biology. In addition to upholding the GenBank nucleic acid sequence database, which accepts information over an international partnership with the DNA Data Bank of Japan (DDBJ) and the European Nucleotide Archive (ENA) as well as from the scientific community, NCBI delivers numerous other types of biological data as well as repossession structures and computational properties for the examination of GenBank and other data.

2.4.2.1 BLAST

Basic Local Alignment Search Tool (BLAST) is the most commonly used tool for manipulating the sequence resemblance. BLAST derives in disparities for usage with diverse query sequences in contrast to diverse databases. BLAST-P program explore protein databases via a protein query.

BLAST utilize the statistical model to yield a bit score and E-value for respectively alignment pair (query to hit). The score gives a sign of how respectable the alignment is and greater score was given for the improved alignment. It is freely available at https://blast.ncbi.nlm.nih.gov/Blast.cgi.

2.5 FUNCTIONAL ANNOTATION

2.5.1 E2P2

The PMN Ensemble Enzyme Prediction Pipeline (E2P2, version 2.1) performed the functional assignment of protein sequences. E2P2 functionally assigns protein sequences by means of homology assignment by participating both single and multiple sequence. The Priam mock-ups of enzymatic function. The reference catalogues utilized in the annotation procedure was depends on Reference Protein Sequence Dataset (RPSD) 2.0. Data for RPSD having experimental support from SwissProt, BRENDA and metacyc.
It is available at https://www.plantcyc.org/e2p2/e2p2-v3.0-ensemble-enzyme-prediction-pipeline-version-3.0

2.5.2 PMN SAVI database refinement

SAVI is a Semi-Automated Validation and Integration pipeline intended to choose the ultimate group of pathways confined in the database for each type. The SAVI pipeline utilize the initial collection of pathways foretold by the Pathologic program for the assumed plant classes, the enzyme forecasts from E2P2 for the plant types, and six sets of pre-sorted SAVI pathways to generate the final set of pathways in each species-confined dataset.

2.6 REGULATORY SEQUENCE ANALYSIS

Even though the straight method of examination of gene is not yet possible, a sum of signals associated to transcription, splicing and translation and are now adequately well considered as to be beneficial in computer estimates of the position and intron exon organisation of genes.

Signal-based approaches search for small sequences that are regularly unvaryingly initiate in and around protein coding regions. These signals signify binding locations of molecules tangled in gene transcription course, in post-transcriptional alterations etc. Models of signals comprise promoter sequences, on behalf of the binding places for DNA polymerase, it encompasses cap-signal, TATA box, CCAAT box and GC box upstream of the gene. Initiation codon ATG, the 'Kozak signal' and Translational signals positioned promptly upstream of preliminary ATG, the 3 stop codons, poly-A signal (AATAAA), splice site consensus and downstream of genes.

2.6.1 Transcription factor prediction

It is clear that transcriptional parameters show a crucial part in the regulation of gene expression in plants. Rigorous analysis of plant mutants has exposed that edifying phenotypes are frequently triggered by alterations in genes for TFs, and a sum of TFs have been recognized that act as important controllers of numerous plant functions. TFs, which control the first step of gene expression, are typically demarcated as proteins containing a DNA-binding domain (DBD) that identify a specific DNA sequence. In addition, proteins lacking a DBD, which interrelate with a DNA-binding protein to form a transcriptional complex, are often characterized as TFs.

A huge sum of transcription factors, predominantly from tomato have been revealed to disturb carotenoid build-up over the control of fruit ripening. Tomato RIN encodes a *MADS box* transcription factor and signifies a worldwide chief controller of fruit ripening. Kang *et al.* (2017) reported that ripening inhibitor (RIN) and RAP2.2 transcription factors were described to regulate PSY and PDS expression in tomato and leaves of Arabidopsis. This specifies that transcription

factors tangled in carotenoid genes might occur in cassava for the transcriptional regulation of carotenoid genes.

2.6.1.2 Transcription factor database (TFDB)

The Plant Transcription Factor Database is an unsegregated database that delivers evidently whole sets of transcription factors (TFs) and other transcriptional regulators (TRs) in plant types whose genomes have been totally sequenced and annotated. PlantTFDB 2.0 covers TFs from 49 species incorporating the key ancestries of the plant kingdom,1 from moss, 9 from green algae, 3 from gymnosperm, 1 from fern, and 35 from angiosperm. Using the advanced pipeline, 53 319 TF recognized from 49 species and grouped into 58 families. They made both computational annotation and physical curation for those assumed TFs (Jin *et al.*, 2017). It is available at http://planttfdb.cbi.pku.edu.cn/.

2.6.1.3 JASPAR

JASPAR is an open database of annotated, high superiority, matrix-based transcription factor binding site outlines for multicellular eukaryotes. The outlines resulting wholly from circles of nucleotide sequences experimentally established to bind transcription factors. The database completed by a web interface for browsing, exploring and subset selection, an online sequence examination utility and a collection of programming gears for genome-wide and comparative genomic investigation of regulatory regions. JASPAR is available at http://jaspar.cgb.ki.se (Sandelin *et al.*, 2004).

2.6.2 CpG Islands

Bird *et al.* (1987) reported that CpG islands are DNA sections that comprise an elevated occurrence of CpG dinucleotides comparative to their incidence in the majority genome, in which a CpG doublet happens at only one fifth of the frequency that would be predictable from the G + C content. The presence of CpG islands in plants was proposed by the outcomes of analyses of DNA sequences of plant genes. Ashikawa *et al.* (2001) proposed that CpG-rich clusters in plants with short genomes may be beneficial for recognizing genes that is expressed within the genomic DNA.

In plants with small genomes, such as rice and A. thaliana, a big amount of the CpG islands were related with genes, which recommends that plant CpG islands may also be valuable for recognizing unidentified genes in different sequences of the genome (Dunham *et al.*, 1999).

It is available at https://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot/.

2.6.3 miRNA analysis

A microRNA is a small non-coding RNA molecule (containing about 22 nucleotides) seen in plants, animals and some viruses, that facilitate in RNA silencing and post-transcriptional control of gene expression. Present researches have exposed that miRNA harmfully delimited their specific gene expression at both post-transcriptional and post translational levels. The results showed that miRNAs organized a countless sum of genes tangled in plant growth and progress, environmental stress reply, signal transduction as well as pathogen attack. (Onsaya *et al.*, 2013).

2.6.3.2 miRNA regulation

In plants, guideline of gene expression by miRNAs is indispensable for usual growth and development, as well as leaf morphogenesis, patterning and polarity launch, developmental timing, floral organ identity, and phytohormone signalling. Moreover, miRNAs are also tangled in plants adaptation to biotic and abiotic stresses. miRBase (a database for miRNA) confined 28,645 mature miRNA entries for plant and animal species from July 2014. These miRNAs were authenticated by means of diverse computational and experimental methods as well as deep sequencing, cloning, northern blots, and real time PCR (Carolina *et al.*, 2013).

2.6.3.3 miRNA prediction tool

2.6.3.3.1 psRNATarget

This server joins hot findings in plant sRNA target acknowledgment, e.g. it differentiates translational and post-transcriptional inhibition, and it hearsays the sum of small RNA/target site pairs that are apparently related with small RNA recognition

action to the target transcript. The psRNATarget was premeditated for high-throughput investigation of next-generation data by executing a dispersed computing pipeline, which runs on a Linux cluster at back-end. The server contributes 3 accessible crossing points to receive user-defined small RNAs and transcript sequences and yields a complete info of small RNAs and alike target positions on contender records laterally with built-in online devices for group downloading, key word probing and the outcomes clarifying (Xinbin *et al.*, 2011).

2.7 CONSERVED DOMAINS

A protein domain is a well-kept-up part of a specified protein sequence and (tertiary) structure that can progress, function, and occur self-sufficiently of the rest of the protein chain. From the discoveries of scientists Frances and Hiroshi, we can say that the conserved domains show a vital role in target gene expression, persuading different phenotype and target gene regulation.

2.7.1 NCBI CD- search

NCBI's border to exploring the CDD with nucleotide or protein query sequences. It utilises a modified of PSI-BLAST, RPS-BLAST to rapidly examine a group of pre-determined site-confined scoring matrices with a protein query. The outcomes of Conserved Domain analysis are obtainable as a functional assignment of protein domains on the user interrogation and can be envisaged as domain multiple sequence alignments with entrenched user queries. High similarity between a query sequence and the CD are displayed as specific hits. The CD-Search assistance delivers extra info, together with data about running CD-Search locally (Bauer, *et al.*, 2015).

2.8 EXPRESSION PROFILES

Gene expression profiles are formulated by the computational analysis through the EST (Expressed Sequence Tag) mining. Through this profiling, each gene expressed in different parts of the plant can be elucidated.

2.8.1 EST

Expressed sequence tags (ESTs) are fragments of mRNA sequences resulting via single sequencing responses achieved on arbitrarily designated clones from cDNA libraries. To date, over 45 million ESTs have been produced from over 1400 different species of eukaryotes. For the most part, EST projects are used to moreover counterpart prevailing genome projects or aid as low-cost substitutes for determinations of gene detection. However, with enhancements in accuracy and coverage, they are starting to discover the claim in areas such as phylogenetics, transcript profiling and proteomics (Boguski *et al.*, 1993).

2.8.2 NCBI dbEST

dbEST is a partition of GenBank that comprises sequence data and other info on single-pass cDNA sequences, or Expressed Sequence Tags, from a sum of organisms. dbEST also comprises sequences that are extended than the traditional ESTs, or are formed as single sequences or in small batches. These sequences are the outcomes of diverse display experiments and RACE experiments. The entity that these sequences have in mutual with traditional ESTs, irrespective of length, quality, or quantity, is that there is slight info that annotated in the record (Boguski *et al.*, 1993).

2.9 CHROMOSOME LOCATION

Chromosomal locations are the elucidation of locus of a chromosome, used to make a Synteny map, describes the physical co-localization of genetic loci on the same chromosome within an individual or species. (Mirzaei *et al.*, 2014) done a chromosomal mapping on argonaute genes of several plants including *Arabidopsis thaliana, Brachypodium distachyon*, *Glycine max*, *Medicago truncatula*, *Populus trichocarpa*, and *Vitis vinifera* was plotted using the NCBI map viewer tool and with that result helped for depicting a synteny map and chromosomal location analysis.

2.9.1 NCBI map viewer

The Map Viewer delivers superior browsing abilities for a subsection of creatures in Entrez Genomes. The creature subdivision is publicized underneath and

as well as on Home Page of Map viewer tool. Map Viewer permits to visualize in addition explore a creature's whole genome, presentation of map of the chromosomes, and zoom into gradually larger levels of features, down to the sequence data for an area of interest. The amount and categories of accessible maps differ by creature, and are labelled in the data and search tips case of each creature.

2.10 PHYLOGENETIC ANALYSIS

Phylogenetic analysis has turn out to be a general phase in classification of gene and protein sequences. Still, even with the obtainability of many reasonable and more-or-less instinctive software tools, building of biologically applicable, useful phylogenetic trees remains a method including numerous serious stages that are integrally non-algorithmic, *i.e.*, reliant on choices made by the user. These steps include, but are not inadequate to, set the goals of the phylogenetic analysis, selecting sequences to be examined, and picking approaches hired in sequence alignment building, as well as procedures and constraints utlized to build the real phylogenetic tree (Cvrckova *et al.*, 2016). Outcomes of phylogenetic analysis as sign for (or against) theories on precise evolutionary stories regarding our genes or proteins of interest.

2.10.1 MEGA software

Molecular Evolutionary Genetics Analysis (MEGA) software nowadays implemented to increasingly larger datasets (Kumar *et al.,* 2016). This demanded technical progression of the computation core and the user interface of MEGA. MEGA software functionality has changed to comprise the formation and examination of sequence alignments, the approximation of sequence deviation, the rebuilding and picturing of phylogenetic trees, and the evaluation of molecular evolutionary hypotheses.

Primary stage in MEGA was to recognize a set of homologous sequences and downloading those sequences. In the second step, sequences were associated in MEGA by two diverse algorithms: ClustalW and MUSCLE. Lastly, phylogenetic tree was built from the allied sequences, MEGA offers many diverse approaches. Here we demonstrate the maximum likelihood method, beginning with MEGA's Models feature, which licenses choosing the greatest appropriate replacement model. Lastly,

MEGA delivers a powerful and flexible interface for the final step, really illustrating the tree for publication.

Gene duplications events have added a new functionality in MEGA to mark tree nodes where gene duplications predicted to occur. This system works with or without a species tree. If a species tree was provided, then we able to mark the gene duplications by following Zmasek and Eddy algorithm.

2.11 PATHWAY CONSTRUCTION

The monitoring method of a biological pathway involves recognizing and organizing content, pulling out info physically or by computational knowledge, and accumulating a database by means of suitable software tools. Main stages tangled in building of a biosynthesis pathway includes data pull out, pathway draft, and precise annotated pathway and bring up-to-date pathway.

2.11.1 Information sources

Data sources, which comprise info concerning sequence data, metabolism, signalling, responses, and connections, are a chief basis of info for pathway construction.

2.11.1.1 METACYC

The MetaCyc database (MetaCyc.org) is a easily available complete datasets explaining metabolic pathway information and enzymes from all areas of organisms. Most of the MetaCyc pathways are short-molecule metabolic pathways that is inventively resolved. MetaCyc covers more than 2400 pathways resulting from >46 000 publications, and is the major curated group of metabolic pathways. MetaCyc used by the PathoLogic constituent of the Pathway Tools software as a reference database computationally forecast the metabolic network of any organism that has a sequenced and annotated genome (Caspi *et al.*, 2016).

2.11.1.2 BIOCYC

BioCyc is an assembly of 5700 creature-specific Pathway Genome Databases (PGDBs), each covering the complete genome and foretold metabolic network of

single organism, as well as enzymes, metabolites, metabolic pathways, predicted operons, reactions, pathway-hole fillers and transport systems. The BioCyc propose a diversity of tools for enquiring and examining PGDBs, together with Omics Watchers and utensils for relative study (Caspi *et al.*, 2016).

2.11.1.3 ATIPD

Arabidopsis thaliana Isoprenoid Pathway Database (AtIPD). The database was assembled by means of info on pathways and pathway genes from BioPathAt, KEGG, AraCyc, and SUBA and from the literature. AtIPD can be investigated or browsed to excerpt data and external links connected to isoprenoid pathway replicas, enzyme actions, or subcellular enzyme localizations. To show measurable gene-linked data on curated pathway models, they shaped image annotation and mapping files for combined use with the MapMan tool. Moreover, SBML XML files of the isoprenoid pathway images constructed using the Cell Designer tool. AtIPD therefore denotes a valued source for isoprenoid network study (Vranova *et al.*, 2011).

2.11.2 Formats and standards

Several standard, computationally legible, object-aligned set-ups was established to enable the arrangement, storage, interchange, and describing of pathway information bases and the applicable experimental info. Significant pathway and pathway-connected formats, which are all XML-dependent, consist of, Proteomics Standards Initiative–Molecular Interactions (PSI-MI), Systems Biology Markup Language (SBML) and Biological Pathways eXchange (BioPAX) (Stromback *et al.*, 2006).

2.11.3 Data Integration

Biological pathway data incorporation was intended to slog with sources of data from a diverse origin. As such, two or more databases may not deliver matching information for a given pathway, but assimilating these two databases may produce a more affluent resource for examination. Moreover, the circumstances under which data is composed, either by investigation or by gathering evidence of the issued

material, in any case the subsidiary references show a vital role and attention to the biologists in producing the investigation more eloquent.

Wong, *et al.* (2017) described the presently accessible network data and associated resources for grapevine. With the intention of illustrating data incorporation approaches into network development and analysis in grapevine, berry specific regulators of the phenylpropanoid pathway was found and made a merged network comprising of overlaying maps of co-expression among structural and transcription factor genes, combined with the existence of promoter cis-binding elements, microRNAs, and long non-coding RNAs (lncRNA). The incorporation of all these data in multi-layered networks permits constructing multifaceted maps of molecular regulation and interaction. De Oliveira Dal'Molin, *et al.* (2016) reported that by systems method a genome-scale metabolic reconstruction was developed in union with the use of multi-omics technologies to gain further perceptions into the metabolism of *Setaria italica.* mRNA, protein, and metabolite profusions, were calculated in mature and immature stem/leaf phytomers, and the multi-omics data were combined into the metabolic reconstruction context to seizure crucial metabolic topographies in diverse developmental phases of the plant.

Krishnakumar, *et al.* (2016) described a new tool called ThaleMine, is a complete data hub that assimilates presently included an extensive collection of genomic info of the prototypical plant *Arabidopsis thaliana.*

2.11.4 Pathway building tools

Pathway building tools are required to populate, visualize, and store a pathway. Presently numerous pathway-building tools deliver the facility to excerpt data as well as to assist multiple standard formats., CellDesigner, Cytoscape and JDesigner are graphical settings for making pathways that can import/export SBML models for recreation.

2.11.4.1 Cell illustrator

Cell Illustrator is a software rostrum for Systems Biology that utilize the idea of Petri net for modelling and pretending biopathways. The modern form of Cell

Illustrator 4.0 uses Java Web Start technology and improved with new proficiencies, including automatic graph grid layout algorithms using ontology info; tools using Cell System Markup Language (CSML) 3.0 and Cell System Ontology 3.0. (Nagasaki *et al.*, 2010)

2.11.5 Pathway visualization

Pathway visualization can also be done using cell illustrator but for a elucidated and comprehensive result, we used cytoscape. Cytoscape is an efficient tool helps to convert SBML or XML files to different formats and other image and PDF file and more editing regarding the visualization done using cytoscape.

2.11.5.1 Cytoscape

Cytoscape is an open source software project for assimilating bio molecular collaboration networks with high-throughput expression information and other molecular positions into a combined theoretical framework. Even though applicable to any system of molecular components and connections, Cytoscape is most influential when used in aggregation with huge databases of protein-protein, protein-DNA, and genetic connections that are progressively available for humans and model organisms. Cytoscape's software core delivers basic functionality to outline and enquiry the network; to visualize and assimilate the network with expression profiles, phenotypes, and other molecular states; and to connect the network to databases of functional annotations. The Core is extensible through a straightforward plug-in architecture, permitting rapid development of supplementary computational scrutinizes and features (Shannon *et al.*, 2003).

MATERIALS AND METHODS

3. MATERIALS AND METHODS

This study entitled "*In silico* analysis of carotenoid biosynthesis pathway in cassava (*Manihot esculenta* C.) was conducted in ICAR- Central Tuber Crops Research Institute (ICAR-CTCRI) during 2016-17. Details regarding the experimental materials used and methodology adopted for various experiments are represented in this chapter.

3.1 WORKFLOW FOR THE *IN SILICO* ANALYSIS OF CAROTENOID BIOSYNTHESIS PATHWAY IN CASSAVA

The approaches adopted for the study entitled *in silico* analysis of carotenoid biosynthesis pathway in cassava are discussed here. In the process of sequence retrieval, both nucleotide and protein sequences of template plants were obtained from phytozome and NCBI databases by using keyword and BLAST search. In BLASTp, template plant sequences were given as query against cassava genome for obtaining carotenoid genes in cassava. The obtained gene sequences of cassava were confirmed through functional annotation. E2P2 and conserved domain analysis were used for functional annotation process. This analysis functionally annotated the carotenoid genes in cassava by linking the conserved domains of the template plants and relating the reference pathways in template plants.

Phylogenetic analysis was an attempt to discern the ancestral relationship and the gene duplication events of the carotenoid genes in cassava. This analysis was carried out by MEGA software. In the nucleotide sequence analysis, gene prioritization aims to identify the most promising genes in the pathway. In cassava, the identified genes were prioritized by comparing with the reference pathways. The expression profiling is the process of measuring the activity of genes in plant tissues or other body part in any organism. The expression profile of the carotenoid genes was developed by mining the carotenoid EST datasets in the NCBI EST database. Regulatory sequence is a segment of nucleotide molecule that is capable of increasing or decreasing the expression of specific genes in an organism. TF, miRNA and CpG islands analysis were done for finding the regulatory activities involved in the carotenoid biosynthesis pathway in cassava.

Finally, the data sets collected from different databases such as plant metabolic pathway database, regulatory sequence database, transcriptome database and protein database were integrated in to the proposed carotenoid biosynthesis pathway of cassava. The pathway was constructed and visualized by using cell illustrator (Fig. 1).

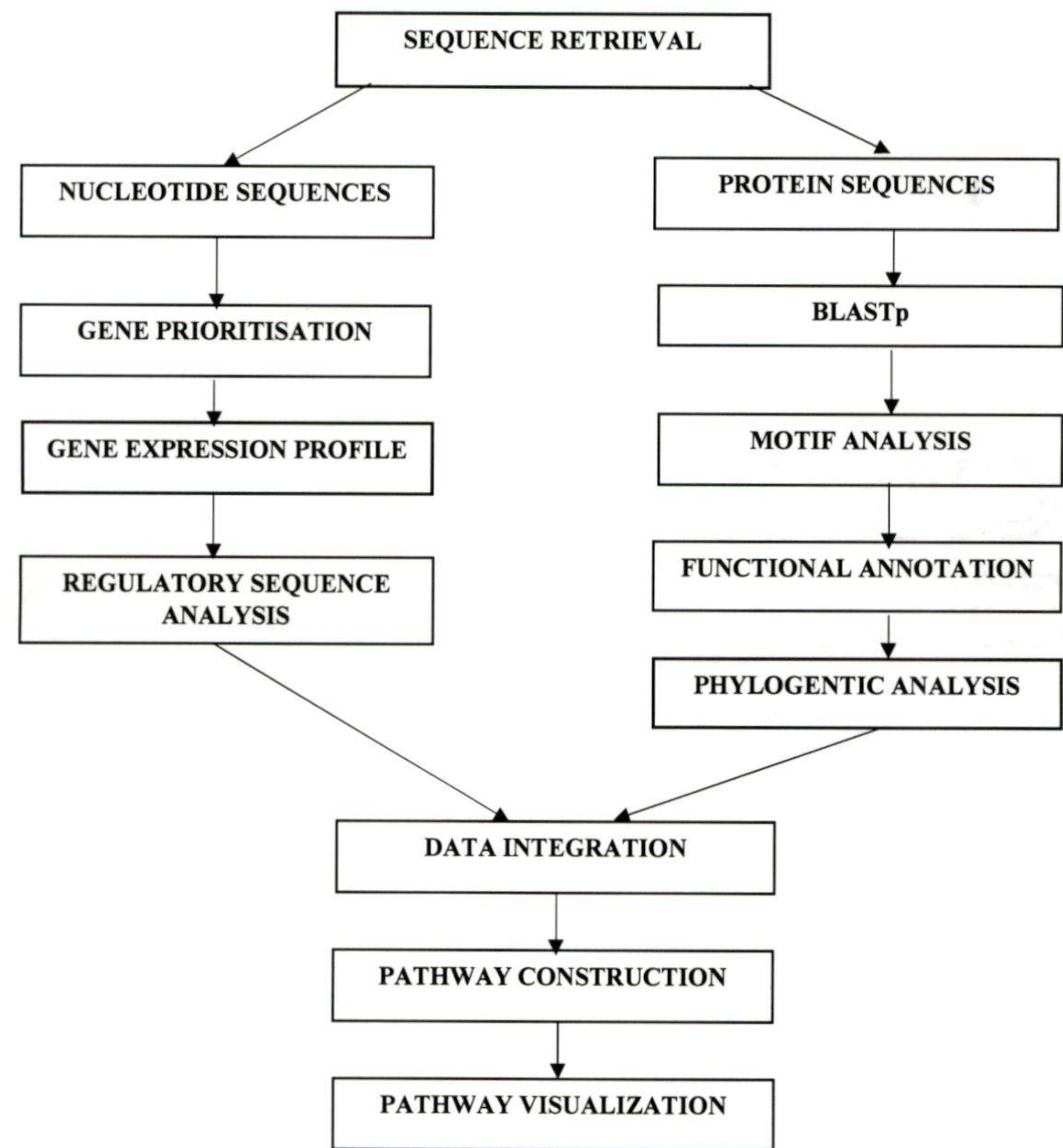


Figure 1. Workflow for the *insilico* analysis of carotenoid biosynthesis pathway in Cassava

3.2 TEMPLATE PLANT SEQUENCE DATA

The preliminary sequence data for the work was obtained from various databases such as NCBI, Phytozome, and Sweet potato garden. Both nucleotide and peptide sequences were used for different analysis and was retrieved from the above-mentioned databases.

3.2.1 Databases

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model plant *Arabidopsis thaliana* . Spud DB (Potato Genomic Resource) comprise sequence files and other related information for the Potato Genome Sequencing Consortium (PGSC) and the carotenoid genes of potato was retrieved from Spud DB. For Tomato, sequences were retrieved from Tomato Genomic Resource Database and database like Sweetpotato GARDEN was used for retrieving sweet potato gene sequences. Phytozome and NCBI were the most common databases used for sequence retrieval. To determine carotenoid genes from template plants, keyword search was done in specific databases. Name of the enzymes involved in carotenoid biosynthesis pathway was used as the keywords. From the query results, genomic and peptide sequences of each plant was downloaded.

3.3 FINDING GENE SEQUNECES OF CASSAVA IN CAROTENOID PATHWAY

Template plant sequences were used as reference sequence for finding the carotenoid genes sequences in cassava. Gbrowse option in Phytozome database was used for locally identifying the gene sequences.

The gene browsers were accessed directly from the Phytozome home page and examinations done in contradiction of one of the genome specific databases. From the search result, the selected gene (or BLAST hit) was contained in the zoomed-in view of the genomic region. Each browser characteristically displayed a gene forecast track. a path of similar or related peptides from connected classes allied in contrary to the genome, secondary EST and one or more VISTA paths identified regions of this genome that was syntenic with other plant genomes comprised in Phytozome.

3.4 FINDING ORTHOLOGS

A Protein BLAST (BLASTp) was done to find the percentage of similarity and syntenic relationship between the gene sequences of cassava and the template plants. BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence database and calculates the statistical significance.

Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Protein sequences retrieved from Phytozome database were inserted to the BLASTp enter query sequence area. BLAST done against to the dataset available in the NCBI database. Alignment of two or more sequence can be done by enabling the align two or more sequence options, so query sequences and target protein sequences was given to the user interface. Hence, result was displayed after running the BLAST program.

3.5 FUNCTIONAL ANNOTATION

Functional annotation was used to assure the integrity of above-mentioned steps. The PMN Ensemble Enzyme Prediction Pipeline (E2P2, version 3.1) used to perform the functional explanation of protein sequences. E2P2 explains protein sequences by means of similarity approach transfer by assimilating both single sequence and multiple sequence (Priam) representations of enzymatic role.

E2P2 was accessed from plant metabolic pathway databases(PMN) currently houses one multi-species reference database called PlantCyc and 76 species/taxon-specific databases, also provides access to manually curated and/or computationally predicted information about enzymes, pathways, and more for individual species. The Ensemble Enzyme Prediction Pipeline (E2P2, version 3.0) annotated protein sequences with Enzyme Function classes comprised of full, four-part Enzyme Commission numbers and MetaCyc reaction identifiers. It is the enzyme annotation pipeline used to generate the species-specific metabolic databases at the Plant Metabolic Network.

E2P2 systematically integrated results from two molecular function annotation algorithms using an ensemble classification scheme. For the carotenoid genes, all

protein sequences were submitted as individual queries against the base-level annotation methods. E2P2 v3.0 used a custom database of annotated protein sequences, which referred to as the Reference Protein Sequence Dataset (RPSD version 3.1). RPSD 3.1 contains approximately 50,182 enzymes and 91,855 non-enzyme sequences, compiled from manually curated or experimentally supported data in UniProt/SwissProt, BRENDA, MetaCyc, and PlantCyc.

3.5.1 Installation:

1. The archived package was retrieved from

   https://dpb.carnegiescience.edu/labs/rhee-lab/software

2. Unzipped and extracted the E2P2 package in the target location:

   tar -xzf e2p2-3.0.tar.gz

3.5.2 Procedure:

E2P2 was built to run on 64-bit Linux systems. As an in-house pipeline, it has not been tested widely on different systems. The input file contained protein sequence data was placed in FASTA format. Paths was specified for input and output files and the program executed with the following command:

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

3.6 REGULATORY SEQUENCE ANALYSIS

Using a variety of bioinformatics tools, significant regulatory analysis and predictions were done in the DNA sequences of cassava. The tools like PlantregMap TFBS (Transcription factor binding site) prediction tool (http://plantregmap.cbi.pku.edu.cn/binding_site_prediction.php), EMBOSS CpG plot (http://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot/), NCBI conserved domain analysis tool(https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi), psRNATarget a Plant Small RNA Target Analysis Server (http://plantgrn.noble.org/psRNATarget/) were used for the analysis of regulatory sequences.

3.6.1 TFBS prediction

For identifying the TFBS, promotor sequences were retrieved from the PlantregMap (http://plantregmap.cbi.pku.edu.cn/download.php) database, the retrieved promotor sequences were arranged according to their respective order of the carotenoid biosynthesis pathway and the sequences were given as the input in the PlantregMap tool, in FASTA format. Threshold p-value was given as per the default parameters (1e-4). After submission of the sequences, the TFBS Results page showed significant TFBS hit for each promotor sequences.

PlantRegMap TFBS tool scanned for TF binding sites from the input sequences. A set of manually curated, non-redundancy and high-quality TF binding Motifs derived from experiments are projected to 156 species using BLAST reciprocal best hits. FIMO (Find Individual Motif Occurrences) was adopted to search TF binding sites in your input sequences for these binding motifs. There are different steps adopted by the PlantregMap sever for analysing the TFBS (Jin *et al.*, 2017).

Initially, the binding motifs were collected from the experiments of Plant Cistrome DB, CIS-BP, JASPAR, UniPROBE, literature and motifs discovered from ChIP-seq peaks using MEME-ChIP. For TFs with more than one motif, PlantregMap tool manually selected the best one preferentially for the motif determined in vivo and presented more similarity with other motifs of this TF. Non-redundant, high-quality binding motif was selected and low-quality TFs were filtered out. Projected set of motifs to other species using the RBHs (Reciprocal Blast Hits) in the same family, and adjusted the motifs using the nucleic acid background in promoter regions.

3.5.2 CpG island prediction

CpG island was predicted by using the online EMBOSS Cpgplot tool (http://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot/). There are mainly three steps involved in this CpG island prediction process.

3.5.2.1 Input sequence

One or more sequences analysed by entering directly into the input sequence form in CpG plot. Sequences can be in GCG, FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP or UniProtKB/Swiss-Prot format. Partially formatted sequences will not be

accepted. The file containing carotenoid protein sequences were uploaded and used as input. Files were saved in the Unix format to avoid hidden windows characters

3.5.2.2 Set parameters

Window Size
This option used to set the parameter for window containing the percentage of CG content and the Observed frequency of CG.
Default value: 100

Minimum length of an island
This option used to set the minimum length of a CpG island
Default value: 200

Minimum observed/expected
This option used to set the minimum average observed to expected ratio of C plus G to CpG in a set of 10 windows that were required before the identification of a CpG island.
Default value: 0.6

Minimum percentage
This option used to set the minimum average percentage of G+C
Default value: 50.

3.5.2.3 Submission

The result was identified by a name called CpG result. This name was associated to the results and appeared in some of the graphical representations of the results.

3.5.3 Conserved domain analysis

NCBI CD tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)was used for analysing the conserved domains in the carotenoid protein sequences of cassava. Carotenoid protein sequence in FASTA format or the GI or Accession of the protein sequence that exists in the Entrez Protein database was given as the query.

Protein sequence queries: There was no length limit on a protein query (CD-Search does not check the length of a protein query sequence).

Nucleotide sequence queries: The maximum length of a nucleotide query was 200,000 base pairs. The examination system decoded into all six reading frames and skimmed the RPS-BLAST databases with the protein products.

3.5.4 miRNA prediction

psRNATarget was used to predict the miRNA targeting the carotenoid genes in the carotenoid biosynthesis pathway of cassava. Three different analysis sections were included in this tool. Submit small RNAs, submit target candidates and submit small RNAs and targets. Submit target candidate was used in this analysis. This section was used to succumb desired contender sequences of attention. A characteristic desirable transcript sequence was nucleotide sequences of carotenoid. The server analysed likely target sites on these succumbed target candidates for (submitted or preloaded) small RNA sequences.

Scoring schema suggested by Axtell was used to score the complementarity between small RNA and the target carotenoid transcript.

3.6 PHYLOGENETIC ANALYSIS

The phylogenetic analysis of the carotenoid sequences retrieved from various model plants and cassava was done by using a software called MEGA (Molecular Evolutionary Genetic Analysis), it is free to download from (http://www.megasoftware.net/home). Following are the steps for constructing a phylogenetic tree through MEGA:

- The acquired and identified sequences were presented on the tree.
- Sequences were aligned by (MSA using ClustalW, T-Coffee, MUSCLE, etc.)
- The constructed tree was estimated by NJ method
- The tree was created and presented

3.6.1 Aligning Sequences

- In MEGA sequence alignment was done by using ClustalW
- Alignment (or refinement) was done in the Analysis Explorer (*Alignment -* > *Open Alignment Explorer* from main menu).

Started with a blank alignment (if the sequences imported from NCBI, or do not have a compatible sequence file).

- With sample sequences in the Alignment Explorer (AE), Alignment option was selected from the menu, and designated the ClustalW icon.
- Alignment parameters was set to defaults value.
- Depending on the length and number of sequences, a progress bar indicated that alignment was running.
- The aligned sequences replaced the previously unaligned sequences in the Alignment Explorer. Exported the aligned sequences to *MEGA* or FASTA format for analysis.

3.6.2 Phylogenetic Tree construction

There were numerous methods for constructing phylogenetic trees from molecular data. These are classified into Distance methods, Parsimony methods, and Likelihood methods.

3.6.2.1 NJ method

NJ method was selected from the dialogue box showing the options and other desired options were selected from the option summery. Options were organized in logical sections. Phylogeny Test and Choices was selected to evaluate the dependability of a phylogenetic tree, MEGA delivers the Bootstrap assessment. This assessment practices the bootstrap re-sampling plan, so the number of replicates were entered. Appropriate examinations and the phylogeny implication method was allowed for given data set.

3.6.3 Visualizing and exploring data and result

The current session was saved and the export tree option was used to export the tree file in a specified format. The print option was used to print the displayed tree in specific size formats. Finally exit tree option was used to terminate the tree explorer.

3.7 CHROMOSOMAL LOCATION ANALYSIS

Chromosomal position of carotenoid genes of template plants including *Arabidopsis thaliana, Solanum lycopersicum, Solanum tuberosum, Populus trichocarpa,* and *Manihot esculenta* was plotted using the NCBI map viewer tool

Following are the steps used for locating the chromosomal positions of sample genes:

3.7.1 Constructing queries

The search bar near the top of the page was used to enter the name of carotenoid genes. The chromosome search field was optional, as described below.
The types of terms that was used for searching vary by organism. Some types of terms were generally searchable in all genomes.

3.7.2 Search output: genome view

The results of a genome-wide search included two main items:

1. A illustration of all the chromosomes (to scale), with red tick inscriptions displayed the site of carotenoid genes that confined the exploration term
2. A horizontal summary of markers that confined the exploration term, and the chromosome and exact maps on which those genes were found.

Under each chromosome explicit, the chromosome sum was publicized in blue. The sum of knockouts on a chromosome was designated in red, underneath the chromosome serials. When the exploration period happens on a sequence with unidentified chromosomal site, the sum of matches to the span was designated by "not placed."

3.7.3 Search output: Map View

The Download/View Sequence/Evidence link displayed region-specific functions, namely downloading the genomic sequence.

3.8 EXPRESSION PROFILE OF CAROTENOID GENES

Transcript levels of *Manihot esculenta* carotenoid genes were analysed by EST (Expressed sequence Tag) mining. Initially, EST mining was performed in the NCBI EST database (http://www.ncbi.nlm.nih.gov/dbEST/) using the blast tool, following are the steps executed for generating expression profile of carotenoid genes in cassava:

3.8.1 Collection of carotenoid gene sequences

The nucleotide sequences of the genes were retrieved from the Phytozome database (https://phytozome.jgi.doe.gov/pz/portal.html#).

3.8.2 Query input

The query sequences were used for a BLAST search and pasted in the 'Search' text area.

3.8.3 Choose search set

BLAST database was designated as of the normal list by means of the pull-down option and expressed sequence tags (EST) was chosen as the database from the menu. In this analysis, the name was given as *Manihot esculenta* in the organism field.

3.8.4 Program selection

The default megablast option was selected. Clicked the "BLAST" button and submitted the search to BLAST server for processing. When processing completed, result was displayed automatically.

3.9 PATHWAY CONSTRUCTION

Cell illustrator was used to construct the carotenoid biosynthesis pathway of cassava.

3.9.1 Cell Illustrator Desktop Start-up

Cell Illustrator is a Java desktop application. After completion of the CI installation, short cuts to the cell Illustrator start-up script appeared on the desktop folder, in the start Menu and/or in the Applications folder.

To start Cell Illustrator, one of the following steps were used:

Cell illustrator was executed by locating the short cut from the above-mentioned locations

Cell illustrator was executed by locating the start-up script, ci50.bat or ci50.sh from the installation folder

Located the Cell Illustrator jar file ci-application.jar in the installation folder and executed it by double clicking it.

3.9.2 Invoking Cell illustrator and constructing carotenoid pathway

Biosynthesis and enzymatic responses was significant events in organisms. A model of a biosynthesis and enzymatic reactions were created easily by using association connectors. To create a model, which involves biosynthesis and enzymatic reactions, executed the following steps.

When starting the cell illustrator, first step is the creation of new canvas, following are the steps to create a new canvas.
1. First, create new canvas icon was selected.
2. Select the new menu option and save it by a suitable name.

3.9.3 Adding Elements

3.9.3.1 Adding an Entity/a Process

Following steps were used for the addition of entity/process into the canvas.

1. Icons on behalf of objects and procedures were selected after the Topmost Toolbar.

2. Clicked and selected the insert Entity/Insert Process.

3. Selected Insert Entity from Edit in the Menu Bar. The new element was introduced into the centre of the work.

4. New element was added by selecting the biological element from the dialog box

3.9.3.2 Adding a Connector

Following steps were used for connecting and entity and a process

1. Selected one of the connector icons from the Top Toolbar,
2. Selected an entity that need to be connected in the canvas.
3. Target process was selected by moving the mouse cursor towards it.

3.9.4 Modifying a model

3.9.4.1 Modifying an element

Selection Mode button was selected from the Top Toolbar.

3.9.4.2 Moving an Element

Element in the canvas was moved by clicking on the component in the work and move it into the novel position.

3.9.4.3 Cutting, Copying, and Pasting an Element

1. clicked the right mouse button and it displayed a popup menu, which allowed cutting, copy, and pasting a component.

3.9.5 Biological element and pathway fragment

Menu options and toolbar buttons were used to add elements. Biological pathways were shaped earlier and further instinctively by dragging and dropping suitable components from the frames that backing construction of pathways.

The Biological Elements frame contained the entity, process and cell component Sheet.

3.9.6 Assigning biological properties to the elements

3.9.6.1 Biological Properties Frame

Biological Properties Frame biological assets and allowed to assess supplementary bio info such as site.

Table 1. HfPNet elements used in cell illustrator for pathway construction

Petri Net Components	Symbol	Biological Equivalent
Discrete Place	○	Discrete Entity A countable biological component or event that is quantified or represented by an integer.
Tokens .	●	Quantity The number of entity items present.
Discrete Transition	▬	Discrete Process A biotic response that changes amounts in separate objects into amounts in other separate objects.
Continuous Place	◎	Continuous Entity A biotic object like a con. of a protein, enzyme, or ion, etc. the amount of which can be signified as a real number.
Universal Place	⊕	Generic Entity A biotic object that is not pre-explained. Cast-off to signify DNA sequence.
Universal transition	⊞	Generic Process Complex Chemical/biological/ physical procedure, e.g., a translation process.
Arc/Connector		
Normal Arc	→	Process Connector Connects input entities to a process and process to output entities.
Inhibitory Arc	⊣	Inhibitory Connector Stops a procedure from being triggered.

3.9.7 Defining model element to an external reference

3.9.7.1 External reference frame

Using this frame, we defining list of references to external databases

3.10. PATHWAY VISUALIZATION

Analysis of assembly and imitation results were approved from exterior of Cell Illustrator, with Cytoscape 3.5.1. The constructed pathway was exported to cytoscape via 'export to cytoscape' option in the cell illustrator software. The exported pathway was visualized in the cytoscape and exported to pdf file by the option export network view as graphics.

RESULTS

4. RESULTS

The main objective of the study was to identify the genes involved in the carotenoid biosynthesis pathways in cassava using Arabidopsis, tomato, potato and sweet potato gene templates, correspondingly construct the carotenoid biosynthesis pathway using the genes identified then analyse the genes involved in carotenoid biosynthesis pathway of cassava using bioinformatics tools and visualize the constructed pathway using Cytoscape. The nucleotide and protein sequences of genes involved in carotenoid biosynthesis pathway of cassava was retrieved from both Phytozome and NCBI databases by executing similarity searches (BLASTp) between predefined carotenoid genes of template plants and cassava genome. The identified nucleotide sequences of respective gene were used for various regulatory sequence analysis like transcription factor prediction by identifying transcription factor binding sites, CpG island analysis, miRNA regulatory analysis and promotor sequence analysis. Expression profile was generated for the genes involved in carotenoid biosynthesis pathway by EST data mining using NCBI EST database. The protein sequences were used for domain/motif analysis, functional annotation and phylogenetic analysis. NCBI map viewer was used to identify the gene position on the chromosome map of cassava. Finally, the carotenoid biosynthesis pathway was constructed using cell illustrator and visualized using cytoscape software. The results obtained in the present study is given below.

4.1 RETRIEVAL OF TEMPLATE PLANT SEQUENCES

The preliminary data set for the work was obtained from the NCBI and the Phytozome database, initially keyword search for all the carotenoid genes were done using NCBI and phytozome database. The results displayed that all template plants have annotated gene sequences except sweet potato. For sweet potato, a database called Sweetpoato GARDEN was used to retrieve the carotenoid gene sequences.

4.2 IDENTIFICATION AND SORTING OF CAROTENOID GENES IN CASSAVA

There were 39 carotenoid genes identified from *Manihot esculenta* by using BLAST and comparative genomic approach. The carotenoid genes length varied from 1320 to 13300 bp nucleotides. Three carotenoid genes (Manes.01G117500,

Manes.16G135300 and Manes.02G018300) had less than 1500 bp nucleotides – such as 1324, 1484 and 1340 bp nucleotides, respectively. Among these carotenoid genes about 16 genes contain 1-5 introns, other 9 members contain 10-25 introns and remaining 14 genes contain 5- 10 introns. In the case of exons about 16 genes contain 1-5 exons, other 11 carotenoid genes harboured 10-27 exons and remaining 12 carotenoid genes contain 5-10 exons. Based on the classification of precursor pathway, core pathway and degradative pathway. A total 39 genes were sorted to different pathways. First 7 genes from DXS to HDR were sorted to MEP pathway, 21 genes from GGPS1 to NSY belongs to the core carotenoid pathway and 11 genes from CCD to AAO3 belongs to carotenoid degradative pathway (Table 2).

4.3 FUNCTIONAL ANNOTATION

All the candidate carotenoid genes were annotated using E2P2 (Ensemble enzyme Prediction pipeline). Protein sequences collected from Phytozome database (https://phytozome.jgi.doe.gov/pz/portal.html) was used for the functional annotation of the carotenoid genes. In analysis, some of the protein sequences of isozymes shows similar function to the protein sequences of other candidate enzymes in the pathway.

The annotated enzymes were associated to reference reactions and pathways in MetaCyc metabolic pathway database (https://metacyc.org/). PRIAM PRofils pour l'Identification Automatique du Métabolisme (ENZYME-SPECIFIC PROFILES for metabolic pathway prediction) (http://priam.prabi.fr/) was a method used for the automated enzyme detection in fully sequenced genome, based on all sequences available in the enzyme database. Genes unable to annotate through E2P2 was functionally annotated by the NCBI-conserved domain analysis tool and results from CDD was combined with the outcomes of E2P2 (Table 3).

Table 2. Genes identified in the carotenoid biosynthesis pathway of cassava

Gene	Name	Location	Length
DXS	Manes.03G017900	Chromosome03:1465554-1472830 reverse	7277
MCT	Manes.08G116100	Chromosome08:28138536-28141510 reverse	2976
DXR	Manes.06G104700	Chromosome06:21452419-21457115 forward	4697
CMK	Manes.02G134400	Chromosome02:9889332-9894740 forward	5409
MDS	Manes.02G017800	Chromosome02:1462869-1466259 reverse	3391
HDS	Manes.04G021700	Chromosome04:2330931-2337998 reverse	7068
HDR	Manes.15G147700	Chromosome15:11538796-11542217	3422
IDI	Manes.13G005300	Chromosome13:689988-693677	3690
GGPS1	Manes.12G098600	Chromosome12:19504527-19506486	1960
GGPS10	Manes.01G117500	Chromosome01:23657031-23658354 reverse	1324
GGR	Manes.11G098400	Chromosome11:16636226-16638742 reverse	2517
PSY	Manes.01G124200	Chromosome01:24153420-24156720	3301
PSY4	Manes.12G139900	Chromosome12:29367478-29371861 reverse	4384
PDS	Manes.16G135300	Chromosome16:28528877-28530360	1484
PDS3	Manes.05G193700	Chromosome05:26661999-26675128	13130
ZISO	Manes.01G001200	Chromosome01:252121-254765	2645
ZDS	Manes.08G128100	Chromosome08:29360970-29368772 reverse	7803
CRITSO	Manes.08G037100	Chromosome08:3396806-3402102	5297
CRITSO2	Manes.07G025600	Chromosome07:2386383-2390737reverse	4355

Table 2. Continued.

Gene	Name	Location	Length
LCYB	Manes.09G008200	Chromosome09:1673410-1676648reverse strand	3239
LEC	Manes.16G099600	Chromosome16:25581180-25586577 reverse	5398
CHY1	Manes.06G152200	Chromosome06:25556654-25558943 reverse	2290
CHY2	Manes.02G018300	Chromosome02:1492088-1493427	1340
LUT5	Manes.09G183600	Chromosome09:29032268-29039339 reverse	7072
LUT1	Manes.09G075500	Chromosome09:11803466-11811199 reverse	7734
CYP97C1	Manes.08G016300	Chromosome08:1520142-1524997	4856
ZEP	Manes.13G124100	Chromosome13:25101072-25107643	6572
VDE	Manes.09G144600	Chromosome09:26239634-26243409 reverse	3776
NSY	Manes.02G068600	Chromosome02:5122509-5125589	3081
CCD	Manes.10G141300	Chromosome10:25249342-25255513	6172
CCD2	Manes.17G071200	Chromosome17:21044733-21047013	2281
NCED3	Manes.03G083500	Chromosome03:13377142-13379442	2301
NCED4	Manes.15G183500	Chromosome15:21175720..21177753	2031
NCED5	Manes.17G071200	Chromosome17:21044840..21046654	2281
NCED6	Manes.03G150400	Chromosome03:24571552-24573342	1791
CCD7	Manes.05G051700	Chromosome05:3869181..3872232	3229
CCD8	Manes.16G075700	Chromosome16:23061820-23065316 reverse	3497
NCED9	Manes.15G102000	Chromosome15:7576697..7578484	1788
ABA2	Manes.04G164700	Chromosome04:28544949-28546993 reverse	2045
AAO3	Manes.04G051100	Chromosome04:8019081..8027427	8389

Table 3. Functional annotation of carotenoid genes by E2P2

Gene ID	BLAST	PRIAM	METACYC reaction ID
Manes.03G017900	EF00121	EF00121	DXSPREDISOM-RXN 1.1.1.267
Manes.06G104700	EF01145	EF01145	DXS-RXN 2.2.1.7
Manes.06G152200	EF11705	EF05788	DXS-RXNIF 1.5.2
Manes.09G075500	EF11505	EF11505	RXN-8671
Manes.09G075500	EF07680	EF07680	RXN-12226
Manes.08G016300	EF07680	EF07680	RXN-12226
Manes.12G139900	EF01529	EF01529	RXN-13323
Manes.09G144600	EF05771	EF05771	RXN-7984
Manes.04G051100	EF00653	EF00653	Aldehyde-oxidase- RXN
Manes.17G071200	EF00343	EF00343	RXN-698
Manes.15G102000	EF00343	EF00343	RXN-7973
Manes.05G051700	EF07535	EF07535	RXN-13642
Manes.16G081600	EF07647	EF07647	RXN-8180

4.4 TFBS PREDICTION

Study of the promoter gene sequences of 39 genes in cassava revealed that more than 39 different motifs were involved in the action of TFs (Fig. 2). The maximum number of TF motifs and their recurrences was detected for BBR-BPC followed by DOF transcription factor families. A alike sum of TF binding sites were found in the promoter sequence of many genes. Among the recognized TF families, highest number of TFs were related to genomic regulation (29%), growth and development (24%) (Fig. 5) and hormonal regulation (14%) (Fig. 4), although the TFs that are intricated in the reply to light was about (12%), Abiotic stresses (12%) (Fig. 3) and metabolism (9%) were low. Overall function of TFs was illustrated in (Fig. 6). The major groups of cis-regulatory elements that were recognized, which were associated to hormonal regulation, can be classified into specific phytohormone subgroups, such as abscisic acid (5), cytokinin (0), ethylene (3), gibberellin (6) jasmonic acid (2) and auxin (1) (Table 4).

Table 4. Transcription factors identified in the carotenoid pathway using TFDB

Genes regulated	**Transcription factor**	**Biological process**	**Cellular component**
DXS	BZIP	Regulation of transcription. Response to GA&ABA mediated signalling pathway. Regulation of photomorphogenesis	Nucleus
MCT, ZDS, NSY, AAO3, CGT	DOF	Regulation of transcription. Seed coat development.	Unknown
DXR, NCED9	NIN- LIKE	Response to water deprivation Stomatal movement Response to nitrate	Nucleus
CMK, NCED3	MYB	Regulation of transcription	
ABA2, LYC, HDS, NMT	MIKC-MADS	Response to cold Response to Gibberellin Positive regulation of floral development	Nucleus Cytoplasm
HDR, AP2	GRAS	Response to ethylene, Jasmonic acid and ABA. Negative regulation of GA mediated signalling pathway. Negative regulation of seed germination.	Nucleus
ZCO, VDE, LUT5, IDI	AP2	Regulation of transcription	Unknown
GGPS1, PSY	C2H2	Unknown	Nucleolus
PSY3, PSY4, PDS3, GGPS7, LUT1, NCED2, ADH	BBR-BPC	Regulation of transcription. Response to ethylene Regulation of development.	Nucleus
ZISO, CHY2	TALE	Regulation of transcription	Nucleus
CRITSO	TCP	Regulation of cell size Root development Negative regulation of leaf senescence	Nucleus
NCED6, CRITSO2, CCD7	BHLH	Regulation of Stomatal movement Regulation of cuticle development Regulation of flowering	Nucleus
NCED4, CHY1	ERF	Response to water deprivation Response to ABA Negative regulation of transcription	Nucleus

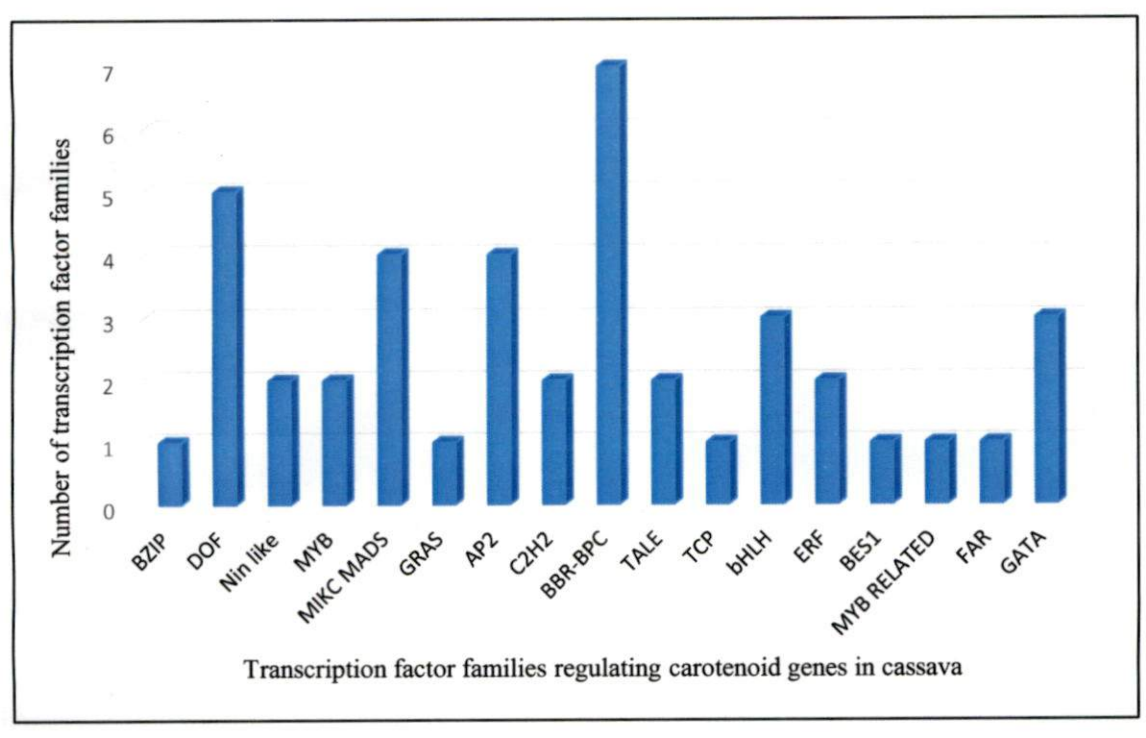


Figure 2. Number of transcription factors regulating carotenoid genes in cassava

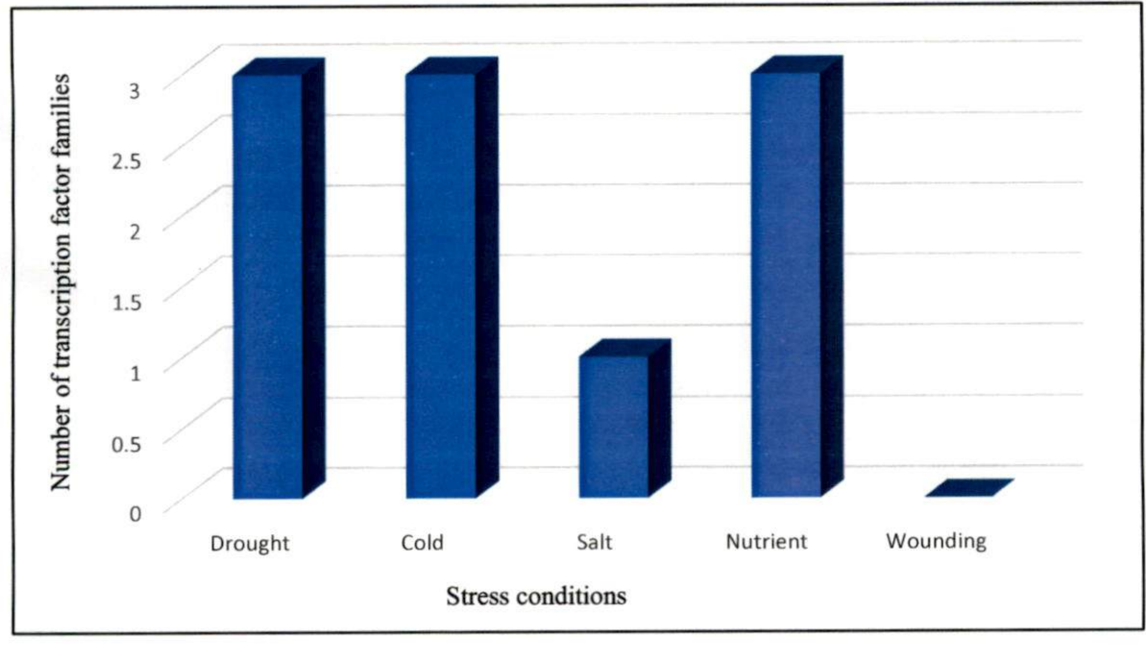


Figure 3. Number of transcription factors related to stress in cassava

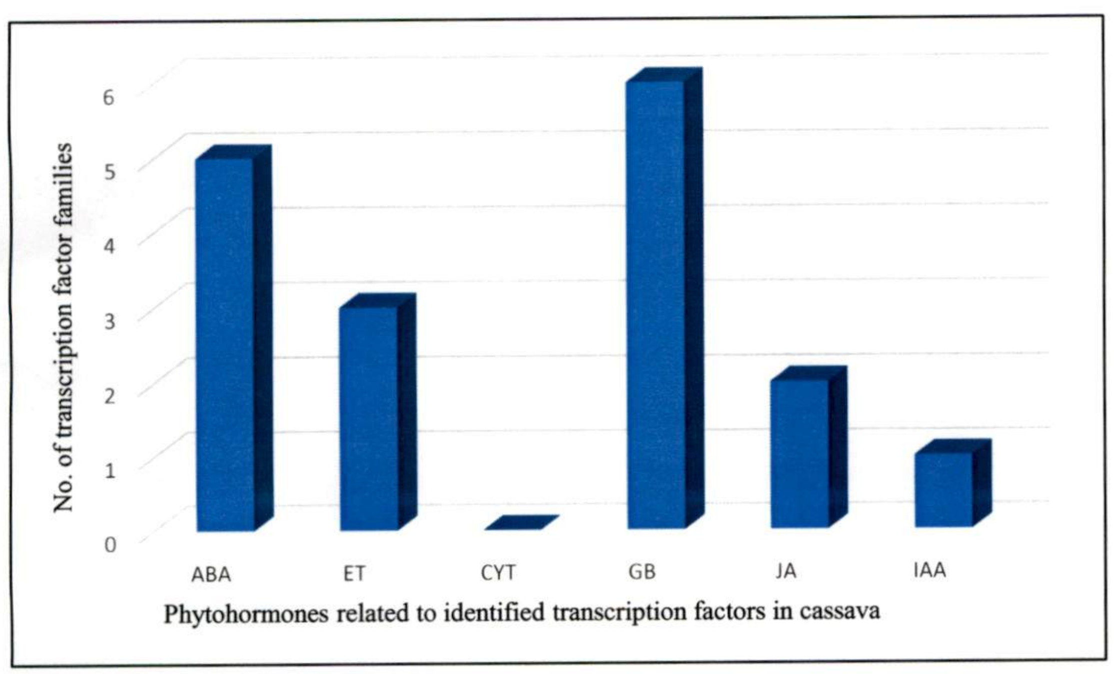


Figure 4. Number of transcription factors related to phytohormones in cassava

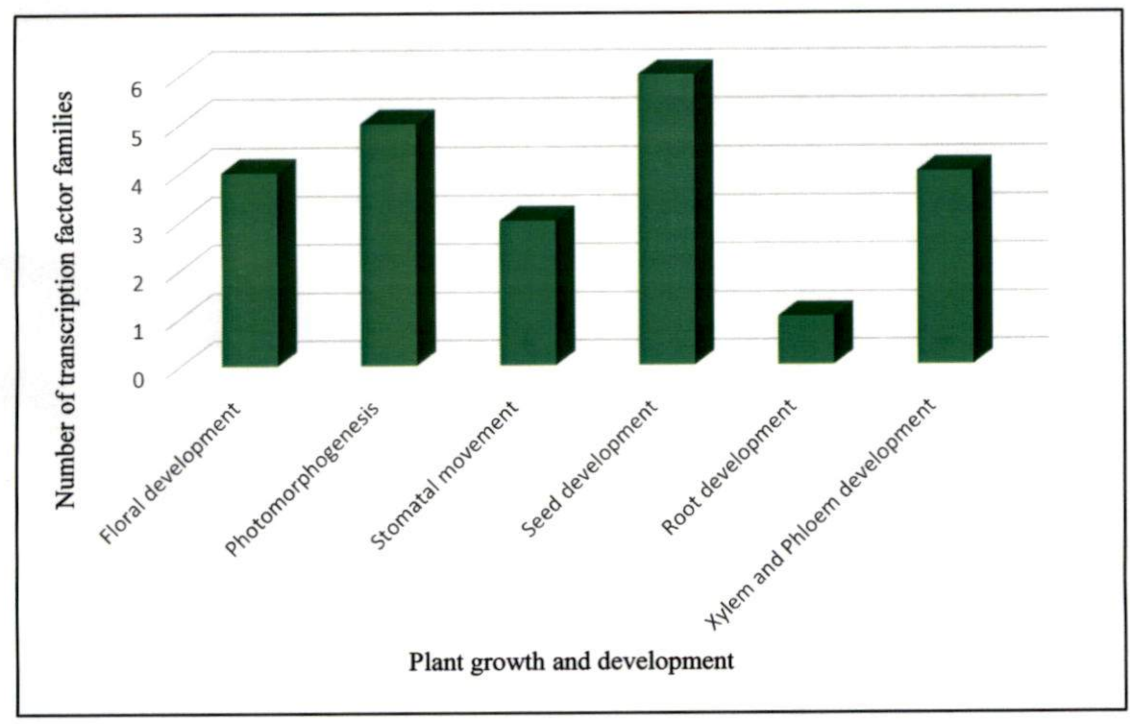


Figure 5. Number of transcription factors related to growth and development of cassava

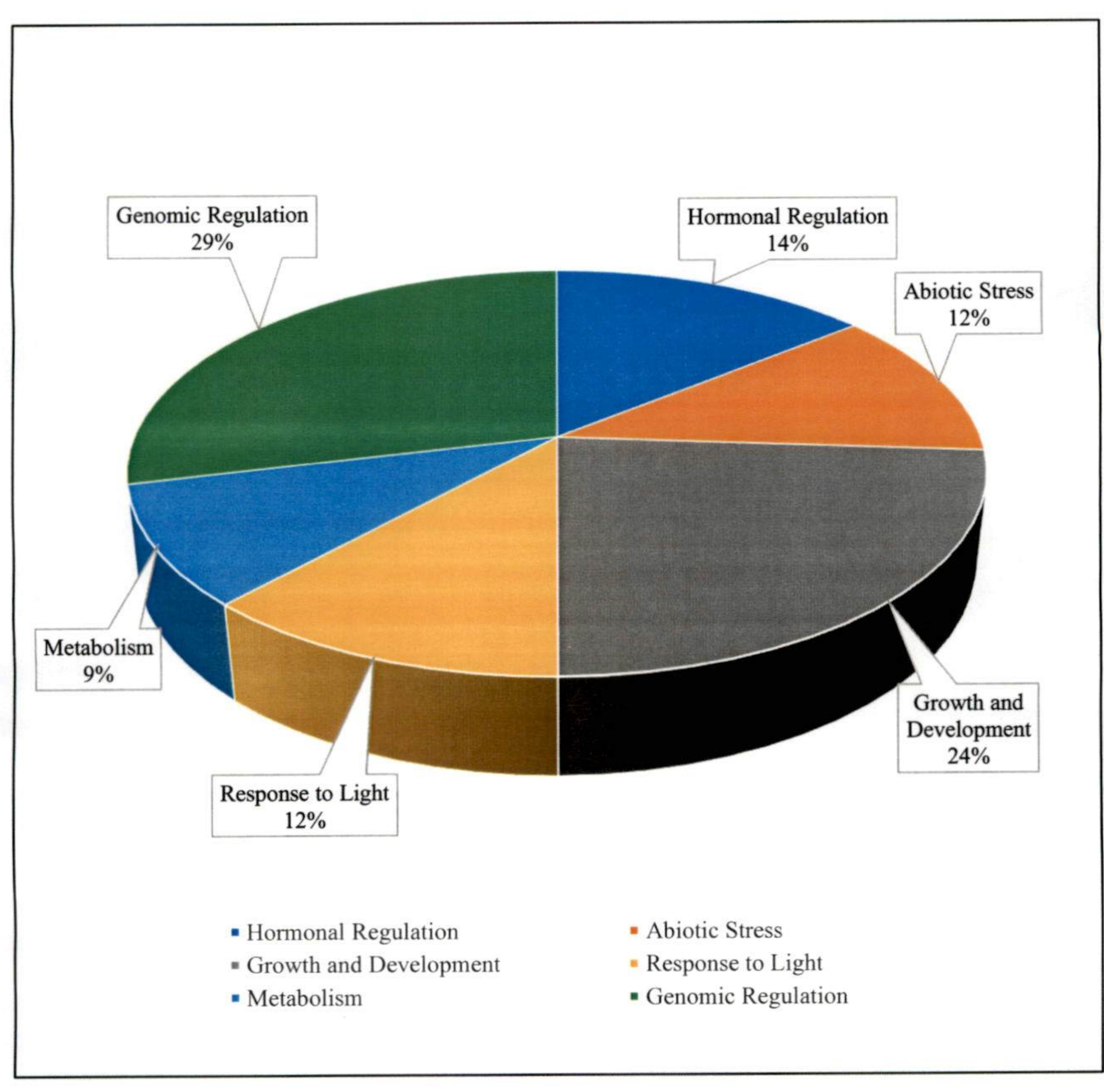


Figure 6. Overall functions of transcription factors regulating carotenoid genes in cassava

Table 4. Continued

Genes regulated	**Transcription factor**	**Biological process**	**Cellular component**
CYP97B3	BES1	Brassinosteroid mediated signalling pathway. Negative regulation of transcription. Seed development. Plant ovule development.	Nucleus Cytosol
CCD8	MYB RELATED	Response to salt stress, GA, ABA, JA and auxin. Response to cadmium ion, photoperiodism and flowering.	Unknown
NCED5	FAR 1	Far red light signalling pathway Positive regulation of circadian rhythm Positive regulation of transcription	Nucleus
LCO, CMGT, CMEGT	GATA	Regulation of transcription	Unknown

4.5 CPG ISLAND PREDICTION

From the carotenoid genes about 19 genes shows CpG islands. Nucleotide sequence of the corresponding genes were used for the CpG island analysis, the analysis was done by using EMBOSS Cpgplot. This online tool can be accessed from (http://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot/).

The ratio of observed to expected number of GC dinucleotides patterns were calculated over a window (sequence region) which was moved along the sequence. The calculated ratios were plotted graphically, together with the regions that match this program's definition of a "CpG island" (a CG dinucleotide rich area). A report file was written giving the input sequence name, CpG island parameters and data on any CpG islands that are found. The output file can be retrieved as two different files, CpG islands graph file and GFF feature format file. The GFF file displayed the length of the sequence and showed the location of CpG island in the respective gene (Table 5).

Table 5. Predicted CpG islands in carotenoid genes using CpG plot

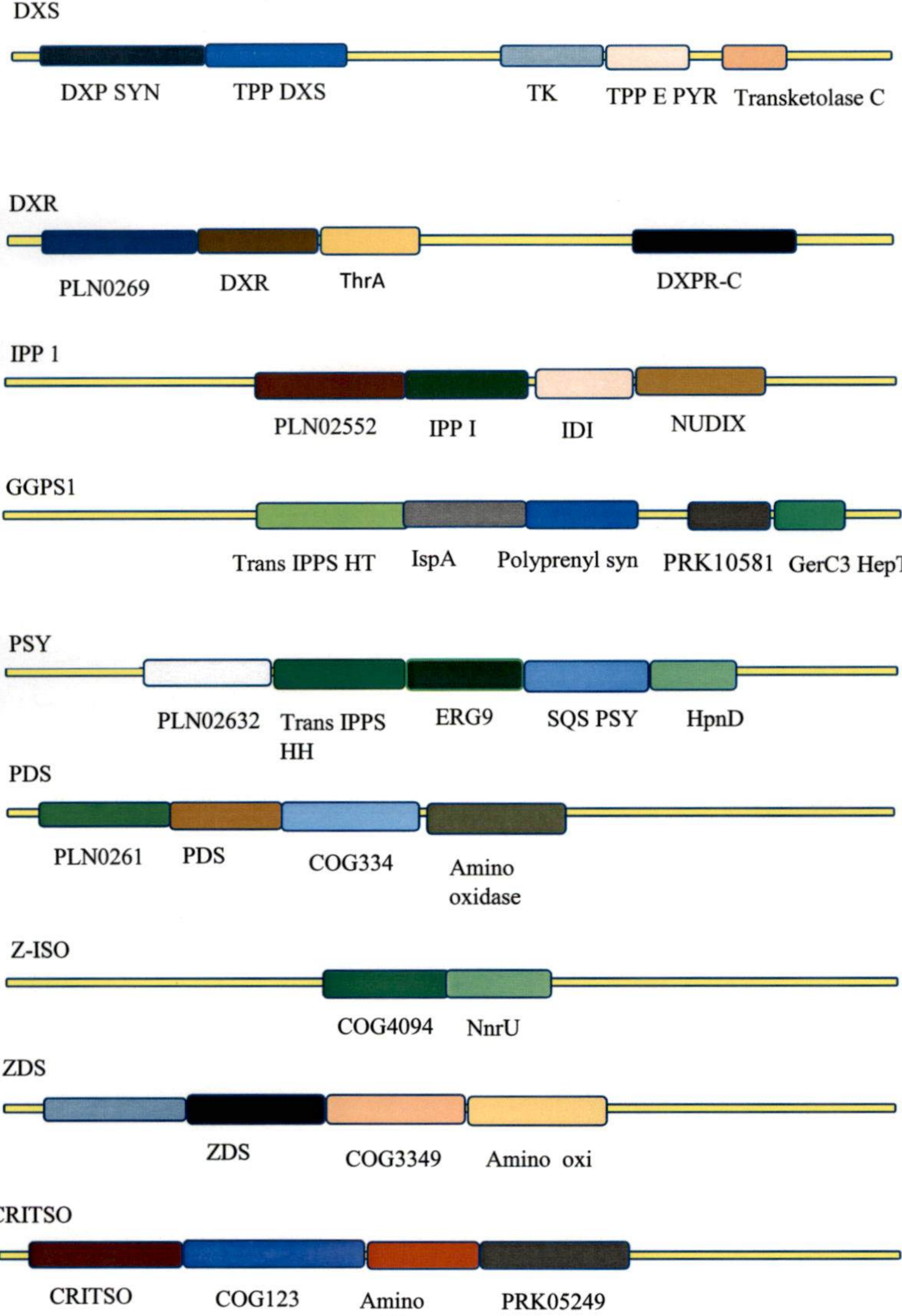
Gene	CpG plot sequence features	ID
Manes.03G017900	Length 100 (100…199) Length 250 (214…463) Length 105(1552...1656)	Manes.03G017900.1.1 Manes.03G017900.1.2 Manes.03G017900.1.3
Manes.15G147700	Length 185 (50…234)	Manes.15G147700.1.1
Manes.12G098600	Length 144 (470…613) Length 254 (631…884)	Manes.12G098600.1.1 Manes.12G098600.1.2
Manes.01G117500	Length 100 (311…410) Length 132 (511…642)	Manes.01G117500.1.1 Manes.01G117500.1.2
Manes.11G098400	Length 189 (303…491)	Manes.11G098400.1.1
Manes.12G139900	Length 148 (147…294) Length 131 (3909…4039)	Manes.12G139900.1.1
Manes.07G025600	Length 124 (1571…1624)	Manes.07G025600.1.1
Manes.09G008200	Length 132 (50…181)	Manes.09G008200.1.1
Manes.16G099600	Length 101 (201…301)	Manes.16G099600.1.1
Manes.06G152200	Length 265 (79…343) Length 104 (413…516)	Manes.06G152200.1.1 Manes.06G152200.1.2
Manes.02G018300	Length 127 (921…1047)	Manes.02G018300.1.1 Manes.02G018300.1.2
Manes.13G124100	Length 111 (351…461)	Manes.13G124100.1.1
Manes.05G051700	Length 170 (207…376) Length 127 (394…520)	Manes.05G051700.1.1 Manes.05G051700.1.2
Manes.16G075700	Length 116 (538…653) Length 155 (963…1117)	Manes.16G075700.1.1 Manes.16G075700.1.2
Manes.17G071200	Length 201 (469…669) Length 199 (858…976)	Manes.17G071200.1.1 Manes.17G071200.1.2
Manes.03G083500	Length 276 (443…718)	Manes.03G083500.1.1
Manes.17G071200	Length 201 (362…562) Length 199 (751…869)	Manes.17G071200
Manes.03G150400	Length 156 (326…481) Length 108 (524…631) Length 108 (667…774)	Manes.03G150400.1.1 Manes.03G150400.1.2 Manes.03G150400.1.3
Manes.15G102000	Length 298 (312…609) Length 166 (621...786) Length 133 (1457…1569)	Manes.15G102000.1.1 Manes.15G102000.1.2 Manes.15G102000.1.3

4.6 CONSERVED DOMAIN ANALYSIS

Bioinformatics analysis of carotenoid protein in cassava and template plants was carried out using the NCBI-Conserved Domains Database and domain sequences were drawn for each gene and placed side by side. Carotenoid proteins usually have PLN02552, PLN02632, PLN02612, PLN02463, PLN02738, PLN02927, PLN02372 and PLN02969 domains and in the investigation, all the participated template plant sequences had PLN02552, PLN02632, PLN02612, PLN02463, PLN02738, PLN02927, PLN02372 and PLN02969 domains but length and location of these domains in each sequence were inconstant. Structural examination of the carotenoid protein sequence in template plants exposed that all of the sequences that had parallel structure but location of domains in the protein was variable; therefore, it appears that all of these proteins have been highly conserved (Fig. 7).

4.6.1 Unusual domains

This analysis showed the information regarding conserved domains in the cassava and template plants. The comparative domain analysis of cassava with other template plants revealed that protein sequences of DXS, CRITSO, LUT1 and ZEP exhibited unusual domains. DXS contain the conserved domains and the unusual domain TPP E PYR, this was organized in between TK and Transketolase C. In CRITSO, there were three unusual domains other than three conserved domains such as NadB, MCRA and COG3380. PLN02648 was the unusual domain present between the protein sequences of LUT1 gene. ZEP holds PRK08274 as the unusual domain (Table 6).


DXS
DXP SYN
TPP DXS
TK
TPP E PYR
Transketolase C
DXR
PLN0269
DXR
ThrA
DXPR-C
IPP 1
PLN02552
IPP I
IDI
NUDIX
GGPS1
Trans IPPS HT
IspA
Polyprenyl syn
PRK10581
GerC3 HepT
PSY
PLN02632
Trans IPPS HH
ERG9
SQS PSY
HpnD
PDS
PLN0261
PDS
COG334
Amino oxidase
Z-ISO
COG4094
NnrU
ZDS
ZDS
COG3349
Amino oxi
CRITSO
CRITSO
COG123
Amino oxidase
PRK05249

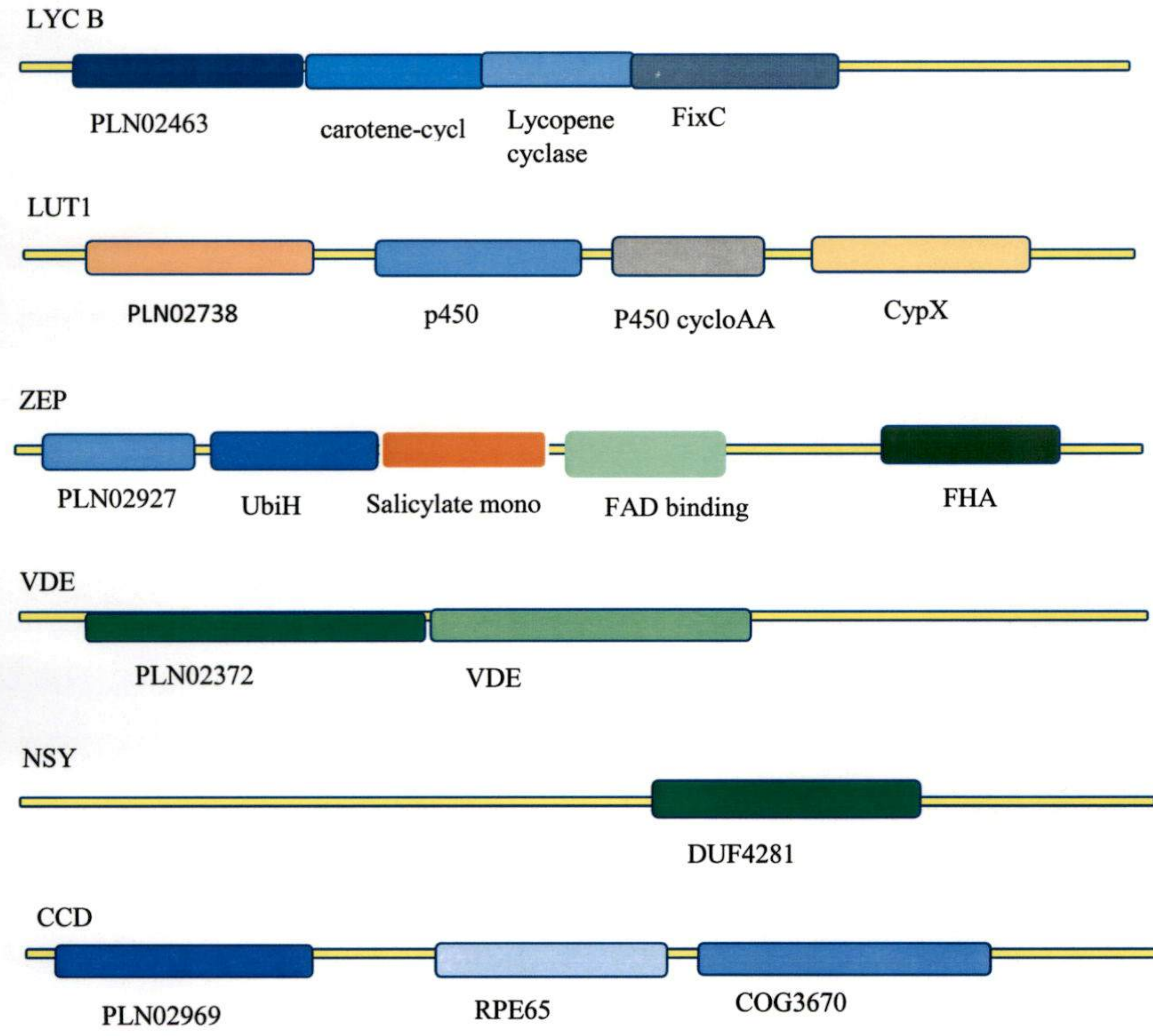


Figure 7. Identified conserved domains of carotenoids in cassava using NCBI-CDD

Table 6. Identified unusual domains of carotenoids in cassava by CDD

Sl.no	Gene	Domain	Functions
1.	DXS	TPP E PYR	Thiamine pyrophosphate (TPP) family, pyrimidine (PYR) binding domain; found in many key metabolic enzymes which use TPP as a cofactor.
2.	CRITSO	NadB, MCRA, COG3380	NadB - Aspartate oxidase (Coenzyme transport and metabolism). MCRA - Myosin-cross-reactive antigen family and may play an important role in the pathogenesis. COG3380 – Predicted NAD/FAD dependent oxidoreductase
11.	LUT1	PLN02648	Allene oxide synthase
12.	ZEP	PRK08274	Tri carballylate dehydrogenase

4.7 miRNA PREDICTION

miRNA is the endogenous non-coding short small RNAs (20–24 nt), or microRNAs (miRNAs) as well as the subdivision of small interfering RNAs (ta-siRNAs). miRNA show a significant part in gene expression regulatory network. For instance, numerous TFs and growth-linked genes have been designated as site of action for these governing small RNAs.

In the miRNA regulatory analysis, miR159a shows a significant regulatory activity towards the MYB domain in CMK and NCED3 genes. miR171b targets the GRAS domain in HDR. LUT, ZCO, VDE and IDI contain AP2 domain and these genes were regulated by the miR172a. miR1027a target CDK domain in ABA2 gene, miR396a target the GF3 domain in LYC gene andmiR164a target the NAC domain in PDS gene in the carotenoid biosynthesis pathway (Table 7).

Table 7. miRNA targeting carotenoid genes and its target sites in cassava

miRNA	Sequence	Target
miR159a	>TTTGGATTGAAGGGAGCTCTA TTTGGATTGAAGGGAGCTCTA	MYB domain (CMK, NCED3)
miR171b	>UUGAGCCGUGCCAAUAUCACG	GRAS TF domain (HDR)
miR172a	>AGAAUCUUGAUGAUGCUGCAU	AP2 factor (LUT, VDE, ZCO, IDI)
miR1027a	>UUUCUAUCUUCUCUUCCAAUC	CDK (ABA2)
miR396a	>UUCCACAGCUUUCUUGAACUG	GF3 (LYC)
miR164a	>UGGAGAAGCAGGGCACGUGCA GGAGAAGCAGGGCACGUGCA	NAC domain (PDS)

4.8 PHYLOGENETIC ANALYSIS

Based on the alignment of full-length sequence of MEP, carotenoid and apocarotenoid proteins, a phylogenetic tree was constructed to determine the evolutionary relationship of these genes in detail. The genes were classified to three groups such as genes from MEP pathway, Core carotenoid pathway and apocarotenoid pathway. The phylogenetic analysis shows that DXS gene considered as the ancestral gene and from which all the descendant genes were formed. The key genes like PSY, MCT, GGR and CCD were initially branched from the DXS gene. PSY gene is the regulatory gene of core carotenoid pathway and CCD gene was one of the candidate gene in apocarotenoid pathway. Other apocarotenoid genes of cassava were gradually diverged and shaped from these key genes (Fig. 8). Comparative phylogenetic analysis of candidate carotenoid proteins of cassava with other template plants shows that PSY gene of cassava and PDS genes of arabidopsis were the ancestral genes and all genes in both cassava as well as other template plants were gradually diverged from these two genes (Fig. 9).

4.9 CHROMOSOMAL LOCATION ANALYSIS

Chromosomal location analysis showed that carotenoid genes were distributed throughout the whole genome. Carotenoid genes were dispersed over 17 out of 18

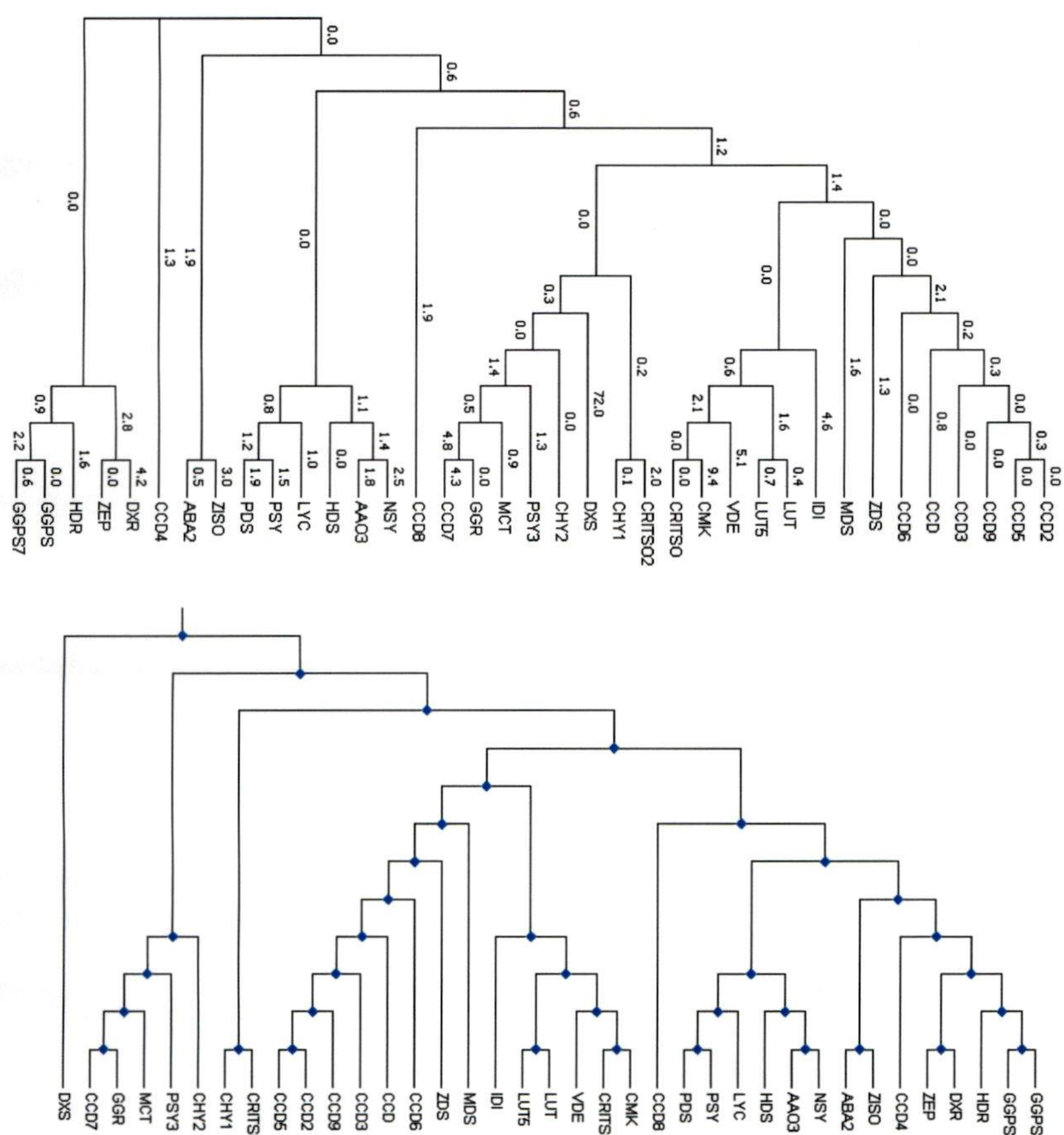


Figure 8. Constructed phylogenetic tree and gene duplications of carotenoids in cassava using MEGA

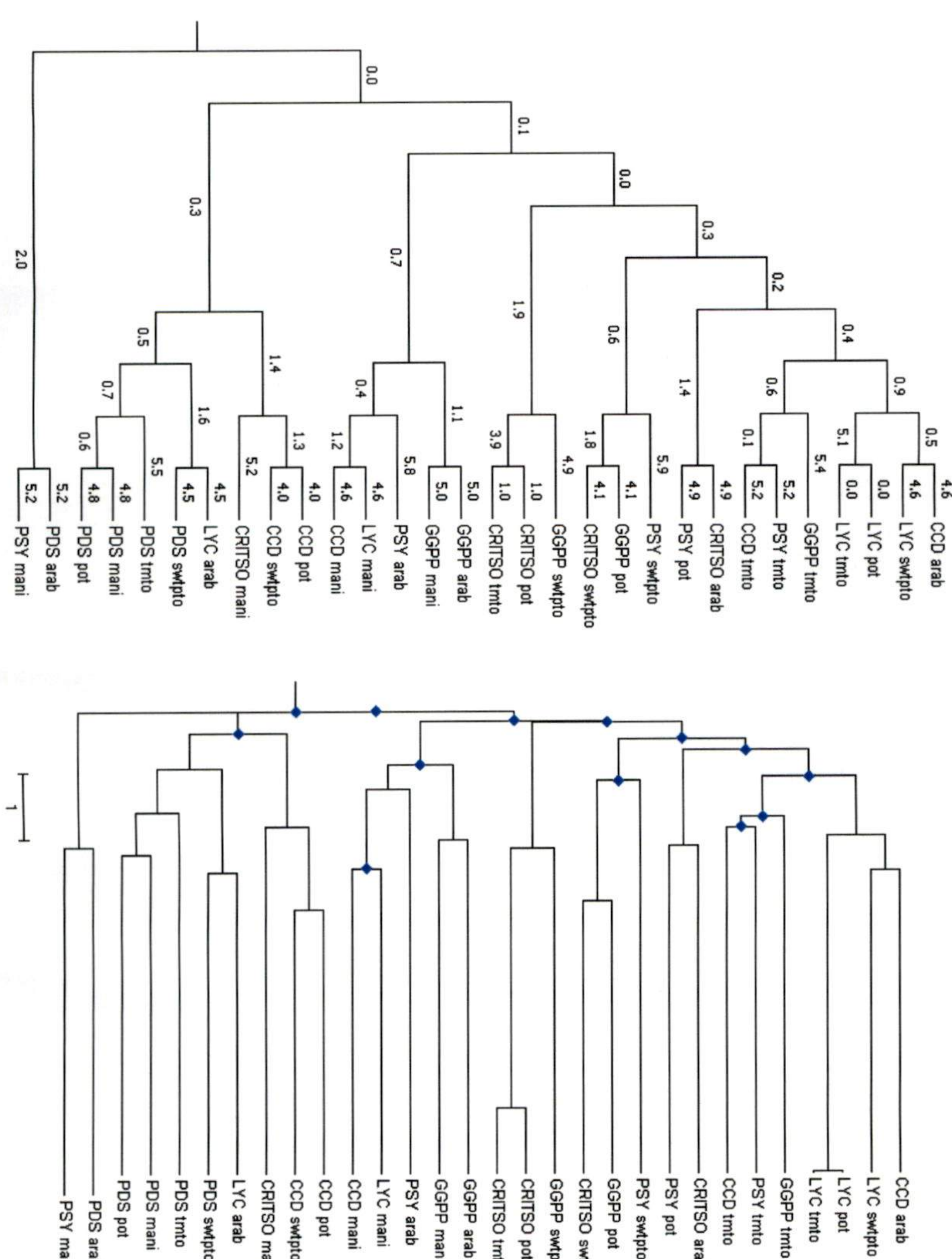
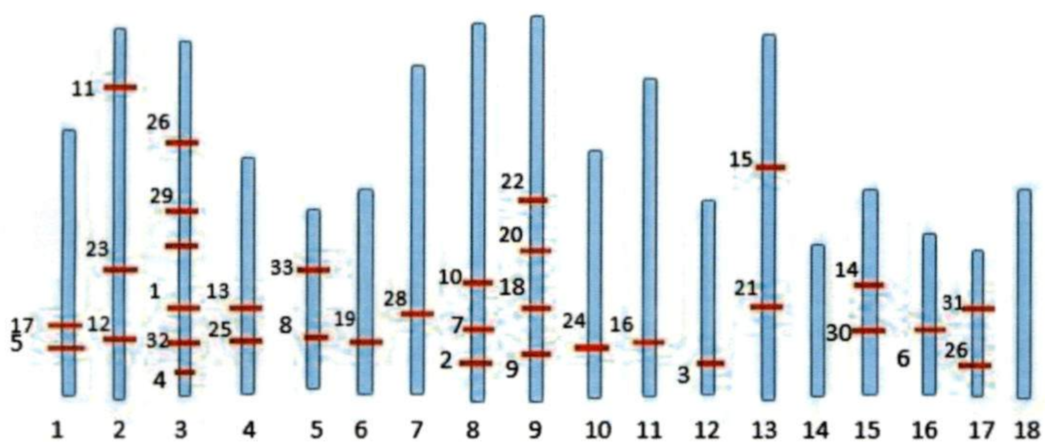


Figure 9. Constructed phylogenetic tree and gene duplications of carotenoids in template plants and cassava using MEGA



1.	DXS	17.	ZISO
2.	DXR	18.	LUT
3.	GGPS1	19.	CHY
4.	PSY	20.	CYP97B3
5.	CRITSO	21.	ZEP
6.	PDS	22.	VDE
7.	ZDS	23.	NSY
8.	CCD1	24.	NCED
9.	LYC	25.	ABA
10.	MCT	26.	AAO3
11.	CMK	27.	CCD2
12.	MDS	28.	CRITSO2
13.	HDS	29.	CCD3
14.	HDR	30.	CCD4
15.	IDI	31.	CCD5
16.	GGR	32.	CCD6
----	------	33.	CCD7

Figure 10. Chromosomal location analysis of carotenoid genes in cassava using NCBI Map viewer

cassava chromosomes. More number of genes were located on chromosome 3 (5 genes), chromosome 9 (4 gene), chromosome 2 and 8 harboured 3 genes each and other chromosomes contain 1-2 genes. No genes were localized on chromosome number 18 (Fig. 10).

5.0 EXPRESSION PROFILE OF CAROTENOID GENES

Expressed sequence tags (EST) data can make available precious data about gene expression research. EST expression analysis in cassava showed that there are many carotenoid genes expressed in different parts of the cassava. GGP1, BCH, CCD7 and CCD4 were the predominantly expressed carotenoid genes in cassava. These genes were expressed in major parts of the plants like stem, root, tuber, petiole and meristem (Fig. 11). This analysis revealed that 23 genes were expressed in the cassava root, 12 genes were involved in the development of stem. Majority of the carotenoid genes in cassava (24 genes) were expressed in leaves, 7 genes were involved in the development of meristem and 4 carotenoid genes expressed in tuber (Table 8).

5.1 PATHWAY CONSTRUCTION

Analysis of carotenoid biosynthesis pathway in cassava was classified to three different segments such as MEP pathway known as the (precursor pathway), core carotenoid pathway and apocarotenoid pathway (carotenoid degradative pathway). Using cell illustrator, the carotenoid biosynthesis in cassava was constructed by interlinking these pathways. Currently no such integrated carotenoid pathway was constructed in cassava. We have constructed a prototype over use of the unknown proofs about the carotenoid biosynthesis in cassava and biochemical knowledge about the reactions. Petri Net component and symbols was used for the construction of the pathway. The precursor molecules were represented by the discrete entity (used to represent biological component) and process connector was used to connect input objects to a procedure and procedure to output entities. Discrete process component was used for representing a biotic response that changes amounts in separate objects to amounts in other separate objects.

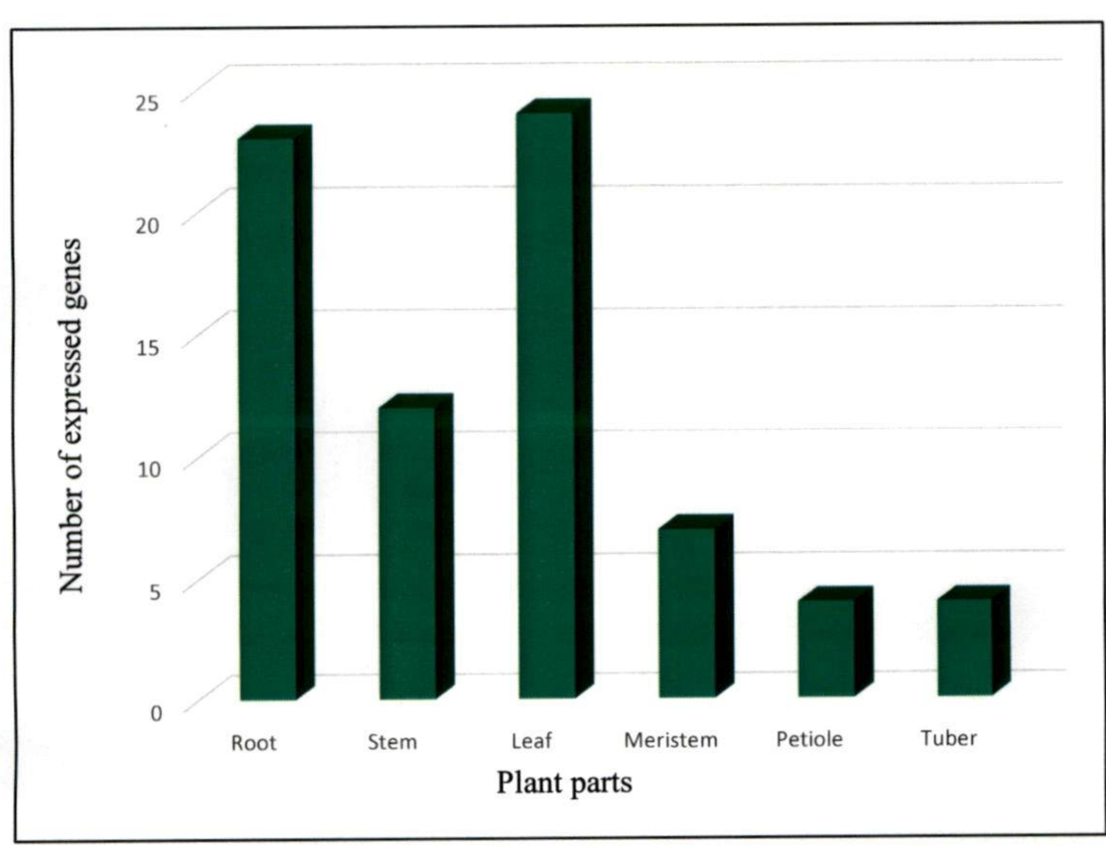


Figure 11. Number of expressed carotenoid genes in cassava retrieved from NCBI EST database

174182

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

Table 8. The expression profile for cassava carotenoid genes from NCBI EST database.

Gene Name	Stem	Root	Leaf	Meristem	Petiole	Tuber
IPP1	✓					
GGP1		✓	✓	✓	✓	✓
GGR		✓	✓			
PSY						
PDS	✓	✓	✓			
ZISO		✓	✓			
ZDS	✓	✓	✓			
CRITSO	✓	✓	✓			
LEC		✓	✓			
BRH		✓	✓	✓	✓	✓
VDE	✓		✓	✓		
NS	✓	✓	✓			
NCED	✓	✓	✓			
AAO3	✓	✓	✓			
ABA2		✓	✓	✓		
CCD1		✓	✓	✓		
CCD3	✓		✓			
CCD4	✓	✓	✓	✓	✓	✓
CCD7		✓	✓	✓	✓	✓
CCD8		✓	✓			
CCD9	✓	✓	✓			
CYP707A1		✓	✓			
CYP707A2	✓	✓	✓			
CYP707A3		✓	✓			
CYP707A4		✓	✓			
CYP711A1		✓	✓			

The black tick mark indicates the expression data for carotenoid genes, and the blank shows that no expression detected

targetp v1.1 prediction results
Number of query sequences: 25
Cleavage site predictions not included.
Using PLANT networks.

Name	Len	cTP	mTP	SP	other	Loc	RC
DXS	720	0.319	0.210	0.055	0.208	C	5
MCT	181	0.073	0.127	0.065	0.765	_	2
DXR	471	0.606	0.039	0.079	0.311	C	4
CMK	388	0.869	0.212	0.038	0.056	C	2
MDS	74	0.271	0.138	0.078	0.714	_	3
HDS	740	0.504	0.063	0.013	0.436	C	5
HDR	462	0.502	0.716	0.033	0.017	M	4
IDI	303	0.942	0.266	0.001	0.020	C	2
GGPS	370	0.906	0.303	0.005	0.015	C	2
GGR	792	0.762	0.117	0.003	0.168	C	3
PSY	429	0.263	0.089	0.087	0.302	_	5
PDS	2681	0.012	0.797	0.012	0.419	M	4
ZISO	372	0.972	0.024	0.008	0.025	C	1
ZDS	586	0.928	0.177	0.007	0.058	C	2
CRITSO	608	0.698	0.347	0.030	0.031	C	4
LYC	504	0.168	0.091	0.056	0.720	_	3
LUT	337	0.059	0.128	0.224	0.877	_	2
CHY	305	0.664	0.123	0.038	0.070	C	3
CYP97B3	2918	0.009	0.831	0.011	0.371	M	3
ZEP	665	0.943	0.102	0.030	0.038	C	1
VDE	483	0.459	0.078	0.059	0.443	C	5
NSY	246	0.257	0.068	0.052	0.287	_	5
CCD	552	0.079	0.043	0.094	0.957	_	1
ABA	277	0.050	0.164	0.239	0.702	_	3
AAO3	391	0.319	0.069	0.026	0.440	_	5
cutoff		0.000	0.000	0.000	0.000		

Figure 12. Screen shot - Subcellular localization of carotenoid genes in cassava using Targetp

Each enzyme in carotenoid pathway was represented and named by discrete entity. The products formed through these enzymatic reactions were represented by the continuous entity, it is used to represent a biological entity like a concentration of proteins, or ions. The biological elements frame was used for representing the cellular localization, miRNA and other regulatory proteins. The pathway model was cross checked with other reference pathways in template plants for avoiding the reaction or enzyme holes in the pathway

5.1.1 Data integration

Subcellular localization results for each enzyme was obtained from TargetP 1.1 server (http://www.cbs.dtu.dk/services/TargetP/) and the result was incorporated to the pathway by specifying the cellular localization of the enzymes in the carotenoid biosynthesis pathway (Fig. 12). Results from miRNA regulatory analysis was integrated to the pathway by representing the miRNA with an inhibitory connector to the corresponding gene encoding enzymes. Information regarding comparative identification of carotenoid genes in cassava was represented by the discrete and continuous entity (Fig. 13).

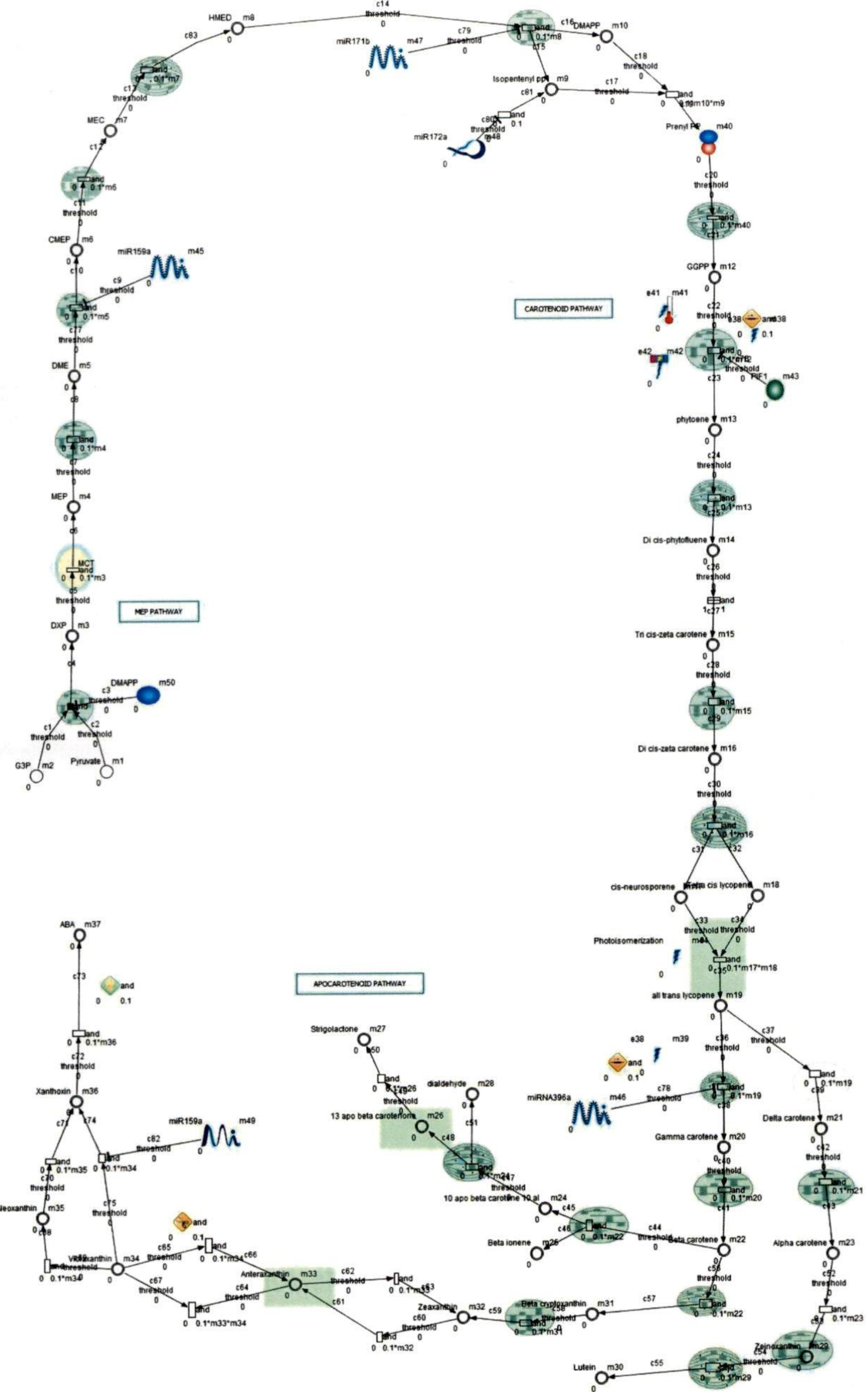


Figure 13. Putative integrated carotenoid biosynthesis pathway in cassava using Cell illustrator

DISCUSSION

5. DISCUSSION

The study entitled "*In silico* analysis of carotenoid biosynthesis pathway in cassava (*Manihot esculenta* C.) was conducted to identify the genes associated with carotenoid biosynthesis as well as to integrate regulatory and pathway information into the putative carotenoid pathway of cassava. The results of this study presented in chapter 4 are discussed here.

Carotenoids constitute a collection of natural pigments formed out of the overall isoprenoid biosynthetic pathway. The reaction basics and enzymes involved in the carotenoid pathway have been well investigated in *Arabidopsis thaliana, Solanum lycopersicum, Solanum tuberosum* and *Ipomea batatas*. Carotenoids are largely synthesized from isopentenyl diphosphate (IPP) and dimethylallyl diphosphate produced by the 2-C-methyl-Derythritol-4-phosphate (MEP) pathway. These compounds have antioxidant property in all organisms and show a vital role in defending cells from injury of radicals such as singlet oxygen. Carotenoids are the chief source of vitamin-A (retinol) in animals and abscisic acid (ABA) in plants. Cassava is one of the most significant staple food crop of the tropics that is fully-fledged under varied environmental circumstances. Due to the absence of cassava genome data, very little is recognized about the genes tangled in the carotenoid biosynthetic pathway of cassava (Cazzonelli *et al*., 2011).

Comparative genomics approaches are valuable to crack complex biological problems (genes and gene networks) that would have been beneficial to develop novel biotechnological tactics in crop improvement programs. Study of the exact biology of agriculturally vital traits administered by multifaceted gene regulatory networks and pathways will allow us to develop smart crops (Kumar *et al*., 2015).

Study of the whole set of carotenoid biosynthetic genes in cassava will provide insight into carotenoid metabolic mechanism in cassava crop. In addition to that, a better understanding of carotenoid biosynthetic genes in cassava will enable the development of conventional and transgenic cassava cultivars with augmented carotenoid levels in the upcoming stages.

Carotenoid genes shows diverse functions in growth and development of plants and animals. They are widely distributed in plant cells throughout the plant world, ranging from small to large plants. They play an important role in animal behaviour, reproduction and survival. Other roles of carotenoids include Retinoic acid for vision, retinol for vitamins, safranal for spices, β-carotene for nutrition, zeaxanthin for antioxidants and astaxanthin for feed additives that play an important role in human health and nutrition. Besides having nutritional roles, carotenoids possess some other subsidiary roles like the ABA release during abiotic stress conditions, mycorradicin comprises in root-mycorrhizal symbiosis, strigol tangled in shoot branching and plant volatiles contain β-ionone. Lutein is an important carotenoid involved in photosystem assembly and light capture (Cazzonelli *et al.*, 2011). In this study, we identified 39 carotenoid genes from *Manihot esculenta* genome by *in silico* analysis which was similar to previous study in Arabidopsis (Ruiz-sola *et al.*, 2012), Tomato (Namitha *et al.*, 2011), Potato (Guliano *et al.*, 2014) and Sweet potato (Kang e*t al.*, 2017) and these studies indicated that carotenoid genes are conservative in plants.

The functional annotation of carotenoid genes by E2P2 and NCBI conserved domain analysis confirms the biological functions of the identified gene sequences. CDD comprises the physically annotated domain mock-ups that enable for protein 3D assembly to improve the domain mock-ups and deliver vision into sequence, structure and function relationships (Marchler-Bauer *et al.,* 2011).

Transcriptional rule of biosynthetic genes also seems to show a dominant part in the regulation of carotenoid making for many other horticultural crops because of the connected variations (Yuan *et al.*, 2015). Study of the promoter sequences of four genes (AtD27, MAX1, 3 and 4) encrypt the proteins that are tangled in strigolactone biosynthesis in arabidopsis and exposed 55 different TFBS. From the TF families that were recognized, most of them were connected to hormonal regulation (26), abiotic stresses (23) and plant growth and progression (19), whereas the TFs that are tangled in response to light (12), biotic stresses (7) and metabolism (6) (Marzec *et al.*, 2015).

The transcriptional regulation analysis in cassava revealed the same result as compared to *Arabidopsis* and tomato. 18 TFs were identified by the TFBS analysis in carotenoid genes of cassava. From this 29% was responsible for genomic regulations, 24% for growth and development, 14% for hormonal regulation, 12% for abiotic stress and light response and only 9% for metabolism.

Plant genome sequences were screened, initially from rice and *Arabidopsis thaliana* for CpG islands and identified DNA segments rich in CpG dinucleotides within these sequences. These CpG-rich clusters seemed in the analysed sequences as separate peaks and arose at frequencies of one per 4.7 kb in rice and one per 4.0 kb in *Arabidopsis*. In rice and *Arabidopsis*, maximum of the CpG-rich clusters was related with genes that recommend that these clusters are beneficial milestones in genome sequences for recognizing genes in plants with small genomes. In disparity, plants with higher genomes, only a limited number of clusters were associated with genes (Ashikawa *et al*., 2001). There was a noticeable connection between the expression of a gene in two or more tissues and the existence of a CpG island in its 5 -end. Among the genes expressed in a single tissue, the genes expressed in callus were separate from those expressed in other tissues in that a large proportion was confined to a class 1 CpG island. These outcomes recommend that plant CpG islands may be valuable for inferring the expression pattern of uncharacterized genes. CpG island analysis of cassava showed significant CpG clusters in 20 carotenoid genes and these genes were expressed in more than 3 plant parts in cassava. CpG island prediction also help us to understand the expression pattern of carotenoid genes in cassava

Evaluation of carotenoid candidates was done based on the identification of domains from the NCBI conserved domain database. Important results of the comparative conserved domain analysis were the identification of unusual domains in the carotenoid protein sequences of cassava. DXS harboured unusual domain like TPP E PYR (Pyrimidine binding domain) thiamine diphosphate (TPP) act as cofactor for TCA, pentose phosphate pathway and isoprenoid pathway. This also showed implicated role in tolerance to DNA damage and activator of disease resistance gene in plants. Protein sequences of CRITSO holds three unusual domains such as NadB, MCRA, COG3380 having different functions like NadB of quinolinate synthase

complex is the key enzyme in denovo synthesis of NAD and it also generate resistance to oxidative stress in plants, MCRA is a domain found in microbes like *Streptococcus pyogenes* and COG3380 is also a domain common in both microbes and plants. LUT1 and ZEP harboured unusual domains such as PLN02648 and PRK08274. These domains are involved in formation of allene oxide synthase and it act as the precursor for defence gene activation by synthesizing Jasmonic acid (Sivasanakar *et al.*, 2000) whereas other domain PRK08274 was involved in the synthesis of tricarballylate dehydrogenase which play a catabolic role in tricarballylate utilization. These domains are present in some microorganisms like *Salmonella enterica.* This indicate that the gene duplication events occur between plants and microbes. Some carotenogenic genes (such as phytoene synthases) have high protein sequence resemblance from bacteria to land plants but some (such as phytoene desaturases, lycopene cyclases, carotenoid hydroxylases and CCOs) have low similarity (Liang *et al.*, 2017)

miRNAs perform a significant part in the direction of various metabolic pathways. For example, apetala 2 (ap2) gene is vital for flower growth and development and signifies one of the principal targets of miR172. Overexpression of miR172 lead to the translational inhibition of the ap2 gene and ap2-like genes and lead to early flowering and other developmental defects. Therefore, clarifying the genetic regulation of carotenoid biosynthetic pathway linking miRNA may offer a stage for operating the carotenoid content in tomato (Kaul *et al.*, 2016). miRNA analysis in carotenoid genes of cassava highlighted different miRNAs like miR159a, miR171b, miR172b, miR1027a, miR396a and miR164a regulating different genes involved in carotenoid biosynthesis pathway in cassava. Kang *et al.* (2017) reported that down-regulation of IbCHY-β by RNA interference (RNAi) resulted in higher levels of β-carotene and total carotenoids as well as salt stress tolerance in cultured transgenic sweet potato cells. So, these results will be useful for metabolic engineering of the carotenoid pathway in cassava.

Priya *et al.* (2014) reported that an analysis was made to explore the phylogenetic relationship among CCD genes (apocarotenoid genes) and to statistically evaluate the sequence conservation and functional separation. In total, 77

genes were recognized from 39 species belonging to 21 families as the outcome of phylogenetic investigation specified the presence of well-conserved subfamilies. Moreover, comparative genomic investigation showed that the gene structures of the CCDs were highly conserved across some different lineage species. The phylogenetic analysis of carotenoid genes in cassava indicate the sequence conservation with the template plants. To study the carotenoid genes in depth, we compared the carotenoid genes and identified 17 pairs of gene duplications among them. These gene duplications were the main driving force of carotenoid gene expansion in cassava.

Chromosomal location analysis revealed that most carotenoid genes are present in 3, 2, 9 and 8th chromosomes of cassava. This result also exposed the gene duplication events of carotenoid genes in cassava. Liu *et al.* (2015) reported that EXPA genes of *Medicago truncatula* showed 17 pairs of gene duplication and these genes were closely located in the same chromosomes that formed the MtEXP clusters. Therefore, these data will be useful for molecular breeding programmes.

Expression analysis of carotenoid genes in cassava disclosed that most genes are expressed in leaf followed by root, stem, meristem, petiole and tuber. Cassava storage root delivers a chief food source for masses of people worldwide. GGPP, BRH, CCD4 and CCD7 were the most commonly expressed carotenoid genes in the plant parts of cassava. These key genes can increase the carotenoid content in tuber and roots of different plants. CCD4 is also a crucial factor of carotenoid/ apocarotenoid content in potato tubers, *Arabidopsis* seeds and senescing leaves and citrus fruits (Giuliano *et al.*, 2014). Incrementing the carotenoid content in storage root of cassava could deliver improved nutritional and health welfares.

By the construction of this pathway, we not only described the pathway as a parts list, but also included the knowledge of the location of the complete pathway, how it is assembled, and whether there exists any regulation of the enzymes or the carotenoids themselves. The genes identified through comparative genome analysis, TF regulation, miRNA regulation, regulation by abiotic stress and subcellular localization of enzymes in carotenoid biosynthesis pathway integrated into the constructed pathway. Pathway was constructed by linking three pathways of

carotenogenesis in plants. The precursor pathway also known as MEP pathway, core carotenoid pathway and apocarotenoid pathway were linked to each other by connecting them with the identified genes. *In silico* approaches are valuable in reducing the complexity regarding the gene networks in cassava or any other plant that will be convenient for developing new biotechnological and bioinformatic tactics in crop improvement programs.

SUMMARY

6. SUMMARY

The study entitled "*In silico* analysis of carotenoid biosynthesis pathway in cassava (*Manihot esculenta* C.*)*" was conducted at the ICAR-Central Tuber Crop Research Institute during 2016-2017. The main objectives of the study were to identify the genes involved in the carotenoid biosynthesis pathways in cassava using templates of *Arabidopsis*, tomato, potato and sweet potato, conduct *in silico* regulatory sequence analysis of carotenoid genes in cassava and to construct and visualize the carotenoid biosynthesis pathway in cassava.

Understanding the complex gene regulatory networks has a greater potential in improving the carotenoid content in staple food crops that could be employed to develop new approaches in biotechnology and bioinformatics for the bio fortification projects. In this study, the main objective was to identify the genes involved in carotenoid biosynthesis pathway in cassava and annotating the identified genes by functional annotation through E2P2. Conserved domain analysis and regulatory sequence analysis were done to obtain the regulatory data regarding the carotenoid pathway in cassava. Phylogenetic analysis, expression profiling and subcellular localization prediction was also done to identify the disbursement of components in the carotenoid pathway. Finally, carotenoid biosynthesis pathway was constructed by integrating the precursor pathway, core carotenoid pathway and apocarotenoid pathway and the data collected from regulatory analysis as well as other *in silico* analysis was incorporated with the constructed carotenoid pathway.

The template plant sequences were retrieved from two different databases, NCBI and Phytozome. BLASTp search was done using the protein sequences of template plants against the carotenoid genome in phytozome and identified 39 carotenoid genes in cassava. Functional annotation revealed the functions of identified carotenoid genes in cassava. This was done by using E2P2 (Ensemble Enzyme Prediction Pipeline) developed by Metacyc and NCBI-CD search. The regulatory sequence analysis such as TF prediction by analysing TFBS using PlantregMap showed 18 significant transcription factors regulating carotenoid genes in cassava, CpG island analysis was done by using the online EMBOSS CpG plot tool

(http://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot/) and the results were useful in analysing the expression pattern of carotenoid genes. miRNA targeting carotenoid genes in cassava were predicted by using the psRNATarget: A Plant Small RNA Target Analysis Server. (http://plantgrn.noble.org/psRNATarget/analysis) identified six significant miRNAs targeting carotenoid genes in cassava. The chromosomal location analysis by NCBI map viewer and the phylogenetic analysis was done by MEGA software that revealed the information regarding conserved genes and gene duplication events of carotenoid genes in cassava. Finally, the carotenoid biosynthesis pathway was constructed by the identified genes and the regulatory elements in cassava.

Study of the complete set of carotenoid biosynthetic genes in cassava, which provide insight into carotenoid metabolic mechanisms in cassava crop and can be utilized for molecular breeding in cassava. In addition, a better understanding of carotenoid biosynthetic genes in cassava facilitate the development of conventional and transgenic cassava cultivars with enriched carotenoid levels in the near future.

REFERENCES

7. REFERENCES

Alder, A., Jamil, M., Marzorati, M., Bruno, M., Vermathen, M., Bigler, Ghisla, S., Bouwmeester, H., Beyer, P., and Al-Babili, S. 2012. The path from b-carotene to carlactone, a strigolactone-like plant hormone. *Science* 335: 1348-1351.

Arango, J., Wust, F., Beyer, P., and Welsch, R. 2010. Characterization of phytoene synthases from cassava and their involvement in abiotic stress-mediated responses. *Planta* 232(5): 1251-1262.

Ashikawa, I. 2001. Gene-associated CpG islands in plants as revealed by analyses of genomic sequences. *Plant J.* 26(6): 617-625.

Bauer, M. A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., Lanczycki C. J., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., and Bryant, S. H. 2015. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* 43: 2-6.

Bird, A. P. 1987. CpG islands as gene markers in the vertebrate nucleus. *Trends Genet.* 3: 342-347.

Burns, A. E., Gleadow, R. M., Zacarias, A. M., Cuambe, C. E., Miller, R. E. and Cavagnaro, T. R. 2012. Variations in the chemical composition of cassava (Manihot esculenta Crantz) leaves and roots as affected by genotypic and environmental variation. *Agric Food Chem.* 60: 4946-56.

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. 1993. dbEST-database for expressed sequence tags. *Nat. Genet.* 4(4): 332-333.

Booker, J., Auldridge, M., Wills, S., McCarty, D., Klee, H., and Leyser, O. 2004. MAX3/CCD7 is a carotenoid cleavage dioxygenase required for the synthesis of a novel plant signaling molecule. *Curr. Biol.* 14: 1232-1238.

Caceres, L. A., Lakshminarayanan, S., Yeung, K. K., McGarvey, B. D., Hannoufa, A., Sumarah, M. W., Benitez, X., and Scott., I. M. 2016. Repellent and

Attractive Effects of α-, β-, and Dihydro-β- Ionone to Generalist and Specialist Herbivores. *J. Chem. Ecol.* 42(2): 107-117.

Carolina, B. T., Germán, P., Sarah, A., Fausto, R. Z., Luis A. B. L., Jorge, D., and Joe T. 2013. Identification of Cassava MicroRNAs under Abiotic Stress. *Int. J. Genomics*. 10(14): 2562-2573.

Carvalho, L. J. C. B., Lippolis, J., Chen, S., Batista de Souza, C. R., Vieira, E. A., and Anderson, J. V. 2012. Characterization of carotenoid-protein complexes and gene expression analysis associated with carotenoid sequestration in pigmented cassava (*Manihot esculenta Crantz*) storage root. *Open Biochem.* 6: 116-130.

Carvalho, L. J., Agustini, M. A. V., Anderson, J. V., Vieira, E. A., Souza, C. R., Chen, S., Schaal, B. A., and Silv, J. P. 2016. Natural variation in expression of genes associated with carotenoid biosynthesis and accumulation in cassava (Manihot esculenta Crantz) storage root. *BMC plant biol.* 16: 133.

Caspi, R., Billington, R., and Ferrer, L. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 44: 471-480.

Cazzonelli, C. I. 2011. Carotenoids in nature: insight from plants and beyond. *Functional Plant Biol*. 38: 833-847.

Cazzonelli, C. I. and Pogson, B. J. 2010. Source to sink: regulation of carotenoid biosyntehsis in plants. *Trends in Plant Sci*. 15: 266-274.

Chen, X., Gao, C., Guo, L., Hu, G., Luo, Q., Liu, J., Nielsen, J., Chen, J., and Liu, L. 2017. DCEO Biotechnology: Tools To Design, Construct, Evaluate, and Optimize the Metabolic Pathway for Biosynthesis of Chemicals. *Chem. Rev.* 45(8): 220-231.

Chen, Y., Li, F., and Wurtzel, E. T. 2010. Isolation and characterization of the Z-ISO gene encoding a missing component of carotenoid biosynthesis in plants. *Plant Physiol.* 153(1): 66-79.

Cvrckova. F. 2016. A plant biologists' guide to phylogenetic analysis of biological macromolecule sequences. *Biologia Plantarum* 60(4): 619-627.

De Oliveira Dal'Molin, C. G., Orellana, C., and Gebbie, L. 2016. Metabolic Reconstruction of *Setaria italica*: A Systems Biology Approach for Integrating Tissue-Specific Omics and Pathway Analysis of Bioenergy Grasses. *Front. in Plant Sci.* 7: 1138-1144.

Deruere, J., Romer, S., d'Harlingue, A., Backhaus, R. A., Kuntz, M., and Camara, B. 1994. Fibril assembly and carotenoid overaccumulation in chromoplasts: a model for supramolecular lipoprotein structures. *Plant Cell* 6: 119-133.

Dogbo, O., Bardat., F, Quennemet, J., and Camara, B. 1987. Metabolism of plastid terpenoids: in vitro inhibition of phytoene synthesis by phenethyl pyrophosphate derivatives. *FEBS Letters* 210: 211-215.

Dong, H., Deng, Y., Mu, J., Lu, Q., Wang, Y., Xu, Y., Chu, C., Chong, K., Lu, C., and Zuo, J. 2007. The Arabidopsis spontaneous cell death 1 gene, encoding a zeta-carotene desaturase essential for carotenoid biosynthesis, is involved in chloroplast development, photoprotection and retrograde signaling. *Cell Res.* 17: 458-470.

Dunham, I., Shimizu, N., Roe, B. A. 1999. The DNA sequence of human chromosome 22. Nature 402: 489-495.

Farmer, E. E. and Mueller, M. J. 2013. ROS-mediated lipid per-oxidation and RES-activated signaling. *Annu. Rev. Plant Biol.* 64: 429-450.

Fester, T. 2002. Occurrence and localization of apocarotenoids in arbuscular mycorrhizal plant roots. *Plant Cell Physiol.* 43: 256-265.

Fester, T. 2002. Stimulation of carotenoid metabolism in arbuscular mycorrhizal roots. *Planta* 216: 148-154.

Fester, T., Wray, V., Nimtz, M., and Strack, D. 2005. Is stimulation of carotenoid biosynthesis in arbuscular mycorrhizal roots a general phenomenon?. *Phytochemistry 66*: 1781-1786.

Floss, D. S., Schliemann, W., Schmidt, J., Strack, D., and Walter, M. H. 2008. RNA Interference-Mediated Repression of *MtCCD1* in Mycorrhizal Roots

of *Medicago truncatula* Causes Accumulation of C_{27} Apocarotenoids, Shedding Light on the Functional Role of CCD1. *Biochemical processes and macromolecular structures* 10: 1104-1109.

Fraser, J., Kiano, M., Truesdale, W., Schuch, P., and Bramley. 1999. Phytoene synthase-2 enzyme activity in tomato does not contribute to carotenoid synthesis in ripening fruit. *Plant Mol. Biol.* 40: 687-698.

Giuliano, G., Al-Babili, S., and Lintig, J. V. 2003. Carotenoid oxygenases: cleave it or leave it. *Trends Plant Sci.* 8: 145-149.

Giuliano, G. 2014. Plant carotenoids: genomics meets multi-gene engineering. *Curr. Opin. Plant Biol.* 19: 111-117.

Gonzalez-Jorge, S., Mehrshahi, P., Magallanes-Lundback, M., Lipka, A. E., Angelovici, R., Gore, M. A., and DellaPenna, D. 2016. ZEAXANTHIN EPOXIDASE Activity Potentiates Carotenoid Degradation in Maturing Seed. *Plant Physiol.* 171(3): 1837-1851.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 11: 78-86.

Hirschberg, J. 2001. Carotenoid biosynthesis in flowering plants. *Curr. Opin. Plant Biol.* 4: 210-218.

Ishikawa, M., Fujiwara, M., Sonoike, K., and Sato N. 2009. Orthogenomics of photosynthetic organisms: bioinformatic and experimental analysis of chloroplast proteins of endosymbiont origin in Arabidopsis and their counterparts in Synechocystis. *Plant Cell Physiol.* 50: 773-788.

Iuchi, S., Kobayashi, M., Taji T., Naramoto, M., Seki, M., Kato, T., Tabata, S., Kakubari, Y., Yamaguchi-Shinozaki, K., and Shinozaki, K. 2001. Regulation of drought tolerance by gene manipulation of 9-cis-epoxycarotenoid dioxygenase, a key enzyme in abscisic acid biosynthesis in Arabidopsis. *Plant J.* 27(4): 325-33.

Jin, J. P., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J. C., and Gao, G. 2017. Plant TFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45: 1040-1045.

Kang, L., Ji, C. Y., Kim, S. H., Ke, Q., Park, S. C., Kim H. S., Lee, H. U., Lee, J. S., Park, W. S., Ahn, M. J., Lee, H. S., Deng, X., and Kwak S. S. 2017. Suppression of the β-carotene hydroxylase gene increases β-carotene content and tolerance to abiotic stress in transgenic sweetpotato plants. *Plant Physiol. Biochem.* 117: 24-33.

Kang, L., Park, S. C., Ji, C. Y., Kim, H. S., Lee, H. S., and Kwak, S. S. 2017. Metabolic engineering of carotenoids in transgenic sweetpotato. *Breed Sci.* 67(1): 27-34.

Kiefer, C., Hessel, S., Lampert, J. M., Vogt, K., Lederer, M. O., Breithaupt, D. E., and Lintig, J. 2001. Identification and characterization of a mammalian enzyme catalyzing the asymmetric oxidative cleavage of provitamin A. *J. Biol. chem.* 276: 14110-14116.

Kaul, A., Yogindran, S., Sharma, D., Kaul, S., Rajam, M. V., and Dhar. M. K. 2016. Carotenoid profiling, in silico analysis and transcript profiling of miRNAs targeting carotenoid biosynthetic pathway genes in different developmental tissues of tomato. *PlantPhysiol. Biochem.* 108: 412-421.

Krishnakumar, V., Contrino, S., Cheng, C., Belyaeva, I., Ferlanti, E. S., Miller, J. R.,Vaughn, M. W., Micklem, G., Town, C. D., and Chan, A. P. 2017. ThaleMine: A Warehouse for Arabidopsis Data Integration and Discovery. *Plant and Cell Physiol.* 58(1): 4-8.

Kumar, A., Pathak, R. K., Gupta, S. M., Gaur, V. G., and Pandey, D. 2015. Systems Biology for Smart Crops and Agricultural Innovation: Filling the Gaps between Genotype and Phenotype for Complex Traits Linked with Robust Agricultural Productivity and Sustainability. *Rev. OMICS A J. Integrative Biol.* 10: 581-601.

Kumar, S., Stecher, G., and Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33(7): 1870-1874.

Lange, B. M. and Ghassemian, M. 2003. Genome organization in Arabidopsis thaliana: a survey for genes involved in isoprenoid and chlorophyll metabolism. *Plant Mol. Biol.* 51: 925-948.

Levonen, A. L., Landar, A., Ramachandran, A., Ceaser, E.K., Dickinson, D. A., Zanoni, G., Morrow, J. D. and Darley-Usmar, V. M. 2004. Cellular mechanisms of redox cell signalling: role of cysteine modification in controlling antioxidant defences in response to electrophilic lipid oxidation products. *Biochem.* 378: 373–382.

Liang, M. H., Zhu, J., and Jiang, J. G. 2017. Carotenoids biosynthesis and cleavage related genes from bacteria to plants. *Crit. Rev. Food Sci. Nutr.* 254(5): 143-157.

Lichtenthaler, H. K. and Buschmann, C. 2001. Chlorophylls and carotenoids - extraction, isolation and purification. *Curr. Protocols in Food Analytical Chem.* 4: 21-26.

Liu, Y., Zhang, J., Li, W., Guo, C., and Shu, Y. 2015. *In silico* identification, phylogeny and expression analysis of expansin superfamily in *Medicago truncatula. Biotech. and biotechnological equipment* 30(1): 197-203.

Luo, Z., Zhang, J., Li, J., Yang, C., Wang, T., Ouyang, B., Li, H., Giovannoni, J., and Ye, Z. 2013. A STAY-GREEN protein SlSGR1 regulates lycopene and beta-

carotene accumulation by interacting directly with SlPSY1 during ripening processes in tomato. New Phytol. 198: 442–452.

Marchler-Bauer, A., Lu, S. and Anderson, J. B. 2011. CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research.* 39: 225-229.

Marzec, M. and Muszynska, A. 2015. *In Silico* Analysis of the Genes Encoding Proteins that Are Involved in the Biosynthesis of the RMS/MAX/D Pathway Revealed New Roles of Strigolactones in Plants. *Int. J. Mol. Sci.* 16(4): 6757-6782.

Mirzaei, K., Bahramnejad, B., Shamsifard, M. H., and Zamani, W. 2014. In silico identification, phylogenetic and bioinformatic analysis of argonaute genes in plants. *Int. J. Genomics*. 55(6): 223-229.

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*. 25: 2-10.

Namitha, K. K., Archana, S. N. and Negi, P. S. 2011. Expression of carotenoid biosynthetic pathway genes and changes in carotenoids during ripening in tomato (*Lycopersicon esculentum*). *Food & Function.* 2: 168–173.

Nishino, H., Murakoshi, M., Li, T., Takemura, M., Kuchide, M., Kanazawa, M., Yang Mou, X., Wada, S., Masuda, M., Ohsaka, Y., Yogosawa, S., Satomi, Y., and Jinno, K. 2002. Carotenoids in cancer chemoprevention. *Cancer and Metastasis Rev*. 21(3): 257-264.

Onsaya, P., Manassawe, L., Punchapat, S., Unchera, V., and Jarunya, N. 2013. Computational Identification of MicroRNAs and Their Targets in Cassava (*Manihot esculenta* Crantz.) *Mol. Biotechnol*. 53: 257-269.

Parkinson, J. and Blaxter, M. Expressed sequence tags: an overview. 2009. *Methods Mol. Biol.* 533: 1-12.

Pogson, B, J. 2008. Plastid signalling to the nucleus and beyond. *Trends Plant Sci.* 13: 602-609.

Priya, R. and Siva, R. 2014. Phylogenetic analysis and evolutionary studies of plant carotenoid cleavage dioxygenase gene. *Gene*. 548(2): 223-233.

Ramel, F., Birtic, S., Ginies, S., Taconnat, L. V., Trinataphylides, C., and Havaux, M. 2011. Carotenoid oxidation products are stress signals that mediate gene responses to singlet oxygen in plants. *Proc. Natl. Acad. USA*. 109(14): 5535-5540.

Rice, P., Longden, I. and Bleasby, A. 2000. EMBOSS: The European molecular biology open software suite. *Trends Genet*. 16: 276–277.

Robson, F., Costa, M. M., Hepworth, S. R., Vizir, I., Pineiro, M., Reeves, P. H., Putterill, J., and Coupland, G. 2001. Functional importance of conserved domains in the flowering-time gene CONSTANS demonstrated by analysis of mutant alleles and transgenic plants. *Plant J.* 28(6): 619-631.

Rodriguez-Concepcion, M. 2010. Supply of precursors for carotenoid biosynthesis in plants. *Arch. Biochem. Biophys.* 504: 118-122.

Rubio, A. 2008. Cytosolic and plastoglobule-targeted carotenoid dioxygenases from Crocus sativus are both involved in b-ionone release. *Biol. Chem*. 283: 24816-24825.

Ruggiero, B., Koiwa, H., Manabe, Y., Quist, T. M., Inan, G., Saccardo, F., Joly, R. J., Hasegawa, P. M., Bressan, R. A., and Maggio, A. 2004. Uncoupling the effects of abscisic acid on plant growth and water relations. Analysis of sto1/nced3, an abscisic acid-deficient but salt stress-tolerant mutant in Arabidopsis. *Plant Physiol*. 136: 3134-3147.

Ruiz-Sola, M. A. and Concepción, M. 2012. Carotenoid biosynthesis in Arabidopsis: a colorful pathway. *The Arabidopsis book/American Soc. Plant Biologists.* 10: 28.

Sandmann G. Evolution of carotene desaturation: the complication of a simple pathway. 2009. *Arch. Biochem. Biophys*. 483: 169-174.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*. 32: 91-94.

Schwartz, S. H., Tan, B. C., Gage, D. A., Zeevaart, J. A., and McCarty, D. R. 1997. Specific oxidative cleavage of carotenoids by VP14 of maize. *Science* 276: 1872-1874.

Shannon, P., Markiel, A., and Ozier, O. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*. 13(11): 2498-2504.

Sharoni, Y. 2012. Carotenoids and apocarotenoids in cellular signaling related to cancer: a review. *Mol. Nutr. Food Res*. 56: 259-269.

Sivasankar, S., Sheldrick, B., and Rothstein, S. J. 2000. Expression of allene oxide synthase determines defense gene activation in tomato. *Plant Physiol.* 122(4): 1335-1342.

Sun, H., Tao, J., Gu, P., Xu, G. and Zhang, Y. 2016. The role of strigolactones in root development. *Plant Signaling & Behavior.*11:345-367.

Song, X. Y.,Zhu, W. J., Tang, R. M., Cai, J. H., Chen, M., and Yang, Q. 2016. Over-expression of *StLCYb* increases β-carotene accumulation in potato tubers. *Plant Biotech. Rep.* 10(2): 95-104.

Straker, C. J., Hilditch, A. J., and Rey, M. E. C. 2010. Arbuscular mycorrhizal fungi associated with cassava (*Manihot esculenta* Crantz) in South Africa. *South African J. Bot.* 76: 102-111.

Stromback, L., Jakoniene, V., Tan, H., and Lambrix, P. 2006. Representing, storing and accessing molecular interaction data: a review of models and tools. *Briefings in Bioinformatics* 7(4): 331-338.

Tan, B. C., Joseph, L. M., Deng, W. T., Liu, L., Li, Q. B., Cline, K., and McCarty, D. R. 2003. Molecular characterization of the Arabidopsis 9-cisepoxycarotenoid dioxygenase gene family. *Plant J.* 35: 44-56.

Tan, B. C., Schwartz, S. H., Zeevaart, J. A., and McCarty, D. R. 1997. Genetic control of abscisic acid biosynthesis in maize. *Proc. Natl. Acad. Sci. USA*. 94: 12235-12240.

Tanaka, Y., Sasaki, N., and Ohmiya, A. 2008. Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. *Plant J.* 54(4): 733-749.

Toledo-Ortiz, G., Huq, E., and Rodriguez-Concepcion, M. 2010. Direct regulation of phytoene synthase gene expression and carotenoid biosynthesis by phytochrome-interacting factors. Proc. Natl. Acad. Sci. U S A 107:11626–11631.

Udoh, L. I., Adesoye, A., and Gedil, M. 2017. Identification and molecular analysis of pro-vitamin A carotenoid genes in cassava (*Manihot esculenta Crantz*), *Mol. Plant Breed.* 8(4): 38-44.

Utama, I. M. 2002. *In vitro* efficacy of plant volatiles for inhibiting the growth of fruit and vegetable decay microorganisms. *J. Agric. Food Chem.* 50: 6371-6377.

Vogel, J. T. 2008. The carotenoid cleavage dioxygenase 1 enzyme has broad substrate specificity, cleaving multiple carotenoids at two different bond positions. *J. Biol. Chem.* 283: 11364-11373.

Vranová, E., Coman, D., and Gruissem, W. 2013. Network analysis of the MVA and MEP pathways for isoprenoid synthesis. *Annu. Rev. Plant Biol.* 64: 665-700.

Vranova, E., Hirsch-Hoffmann, M., and Gruissem, W. 2011. AtIPD: A Curated Database of Arabidopsis Isoprenoid Pathway Models and Genes for Isoprenoid Network Analysis. *Plant Physiol*. 156(4): 1655-1660.

Walter, M. H. and Strack, D. 2011. Carotenoids and their cleavage products: biosynthesis and functions. *Nat. Prod. Rep*. 28: 663-692.

Wei, S. 2011. Enhanced b-ionone emission in Arabidopsis over-expressing AtCCD1 reduces feeding damage *in vivo* by the crucifer flea beetle. *Environ. Entomol*. 40: 1622-1630.

Welsch, R., Maass, D., Voegel, T., Dellapenna, D., and Beyer, P. 2007. Transcription factor RAP2.2 and its interacting partner SINAT2: stable elements in the carotenogenesis of Arabidopsis leaves. Plant *Physiol.* 145:1073–1085.

Wong, D. C. J. and Matus, J. T. 2017. Constructing Integrated Networks for Identifying New Secondary Metabolic Pathway Regulators in Grapevine: Recent Applications and Future Opportunities. *Front. in Plant Sci.* 8: 505-510.

Xinbin, D. and Patrick, X. Z. 2011. psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res.* 39: 155-159.

Yu, Q., Ghisla, S., Hirschberg, J., Mann, V., and Beyer, P. 2011. A new member of the fad red-dependent flavoproteins catalyzing non-redox reactions. *J. Biol. Chem.* 286(10): 8666-8676.

Yuan, H., Zhang, J., Nageswaran, D., and Li, L. 2015. Carotenoid metabolism and regulation in horticultural crops. *Hort. Res.* 15: 1234-1240.

Zybailov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., and Wijk K. J. 2008. Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. PLoS ONE. 3(4): 1994-1999.

“*IN SILICO* ANALYSIS OF CAROTENOID BIOSYNTHESIS PATHWAY IN CASSAVA (*Manihot esculenta* C.)”

By

MUHAMMED SADIQ P. A.

(2012-09-119)

Abstract of Thesis

Submitted in partial fulfilment of the

Requirement for the degree of

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

Faculty of Agriculture
Kerala Agricultural University, Thrissur

DEPARTMENT OF PLANT BIOTECHNOLOGY

COLLEGE OF AGRICULTURE

VELLAYANI, THIRUVANANTHAPURAM - 695 522

KERALA, INDIA

2017

8. ABSTRACT

The project entitled "*In silico* analysis of carotenoid biosynthesis pathway in cassava (*Manihot esculenta* C.)" was conducted at ICAR-CTCRI, Sreekariyam, Thiruvananthapuram during 2016-2017. The main objective of the study was to identify the genes involved in the carotenoid biosynthesis pathways in cassava, construction of the carotenoid biosynthesis pathway in cassava, regulatory sequence analysis of carotenoid genes in cassava and visualization of the constructed carotenoid pathway. The preliminary template data set of carotenoid genes were obtained from NCBI and Phytozome database. 39 carotenoid genes were identified from cassava through comparative genomic analysis and BLAST-p search

Functional annotation of carotenoid genes was done using E2P2 tool and NCBI conserved domain analysis tool. This ensures the function of the carotenoid genes in cassava. The identified genes were used for the regulatory sequence analysis. TF prediction analysis shows that about 18 TF associated with identified carotenoid genes in cassava, CpG island analysis facilitated to understand the carotenoid gene expression pattern in cassava and miRNA analysis showed six significant miRNAs targeting carotenoid genes in cassava. Phylogenetic analysis and chromosomal location analysis revealed the gene duplication events and conserved genes between cassava and the template plants such as *Arabidopsis thaliana, Solanum lycopersicum, Solanum tuberosum, Ipomea batatas and Populus trichocarpa.* Carotenoid biosynthesis pathway in cassava constructed by integrating the identified genes and regulatory elements using cell illustrator software. MEP pathway, core carotenoid pathway and apocarotenoid pathway was interconnected into a single pathway and subcellular localization analysis data was united with this novel pathway. Study of the carotenoid biosynthetic genes and pathways in cassava will provide insight into carotenoid metabolic mechanisms and enrichment of carotenoids in cassava.

103

174182