INTEGRATION OF QUANTITATIVE TRAIT LOCUS (QTL) FOR TUBER COLOUR VARIATIONS WITH GENOMIC INFORMATION IN SWEET POTATO (*Ipomoea batatas* L.)

By

RESHMA T. K.

(2014-09-118)

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

2019

DECLARATION

I, hereby declare that the thesis entitled **"Integration of Quantitative Trait Locus (QTL) for tuber colour variations with genomic information in sweet potato (*Ipomoea batatas* L.)"** 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, associateship, fellowship or other similar title, of any other University or Society.

Place: Vellayani
Date: 29-11-2019

RESHMA T.K.
(2014-09-118)

स्वच्छ भारत
एक कदम स्वच्छता की ओर

भा.कृ.अनु.प- केंद्रीय कन्द फसल अनुसंधान संस्थान
(भारतीय कृषि अनुसंधान परिषद, कृषि और किसान कल्याण मंत्रालय, भारत सरकार)
श्रीकार्यम, तिरुवनन्तपुरम-695 017, केरल, भारत

ICAR- CENTRAL TUBER CROPS RESEARCH INSTITUTE
(Indian Council of Agriculture Research, Ministry of Agriculture and Farmers Welfare, Govt. of India)
Sreekariyam, Thiruvananthapuram-695 017, Kerala, India

ISO 9001:2008

CERTIFICATE

Certified that this thesis entitled **"Integration of Quantitative Trait Locus (QTL) for tuber colour variations with genomic information in sweet potato (*Ipomoea batatas* L.)"** is a record of research work done independently by **Ms. RESHMA T. K. (2014-09-118)** under my guidance and supervision and this has not previously formed the basis for the award of any degree, diploma, fellowship or associateship to her.

Place: Sreekariyam
Date: 29-11-2019

Dr. J. Sreekumar
(Chairman, Advisory Committee)
Principal Scientist (Agrl. Statistics),
Section of Extension and Social Sciences,
ICAR-CTCRI, Sreekariyam
Thiruvanathapuram-695 017

Phone : +91-471-2598551 to 2598554
Director (Per) : +91-471-2598431
(Res) : +91-471-2597211
Sr. Admn. Officer : +91-471-2598193
Fax : +91-471-2590063
E-mail : director.ctcri@icar.gov.in
ctcritvm@gmail.com
Web : http://www.ctcri.org

CERTIFICATE

We, the undersigned members of the advisory committee of Ms. Reshma T. K. (2014-09-118) a candidate for the degree of B.Sc. - M.Sc. (Integrated) Biotechnology, agree that the thesis entitled "**Integration of Quantitative Trait Locus (QTL) for tuber colour variations with genomic information in sweet potato (*Ipomoea batatas* L.)**" may be submitted by Ms. Reshma T. K. in partial fulfillment of the requirement for the degree.

Dr. J. Sreekumar
Chairman, Advisory committee
Principal Scientist (Agrl. Statistics),
Section of Extension and
Social Sciences.
ICAR-CTCRI, Sreekariyam,
Thiruvananthapuram - 695 017

Dr. K. B. Soni
(Member, Advisory Committee)
Professor and Head
Department of Plant Biotechnology
College of Agriculture, Vellayani
Thiruvananthapuram- 695 552

Dr. C. Mohan
(Member, Advisory Committee)
Principal Scientist (Plant Breeding)
Division of Crop improvement
ICAR- CTCRI
Sreekariyam,
Thiruvananthapuram- 695 017

Dr. Swapna Alex
(Member, Advisory Committee)
Professor and Course Director
B. Sc. - M. Sc. (Integrated)
Biotechnology Course
Department of Plant Biotechnology
College of Agriculture, Vellayani
Thiruvananthapuram- 695 552

Dr. M. K. Rajesh
(External Examiner)
Principal Scientist (Biotechnology)
Division of Crop improvement
ICAR- CPCRI
Kasaragod - 671124

ACKNOWLEDGEMENT

With Boundless love and appreciation, I would like to extend my heartfelt gratitude and appreciation to all the people who helped me throughout the project work to make into a reality.

It is with great pleasure, I would like to express my heartfelt gratitude to my most respected advisor Dr. J. Sreekumar, Principal Scientist, ICAR-CTCRI, Thiruvananthapuram for his patience, persistent interest in my work, constant support and encouragement as well as his calm attitude in the difficult situations throughout my study. I am highly indebted to him for his excellent guidance.

My special thanks to the Director of ICAR-CTCRI Dr. Archana Mukherjee and Dr. Sheela Immanuel, Head of the Dept. E.S.S. for allowing me to do my project work.

I take immense pleasure to express my deep sense of gratitude to Dr. Swapna Alex, Dr. K. B. Soni and Dr. C. Mohan, not only for their insightful comments and encouragement, but also for their constructive suggestions that were much helpful throughout my research progress.

My very sincere gratitude towards Dr. Senthil, Scientist and Prakash ettan, Technical Assistant, ICAR-CTCRI who helped me with the technical aspects of my research work including the handling of various instrument. I also would like to thank all the Scientists and Staff of Section of Extension and Social Science, ICAR-CTCRI for their support and help throughoutmy research work.

I am very thankful to Dr. Makeshkumar T., Dr. Sangeetha, Jayakrishnan chettan, Sumayya chechi, Divya chechi, Iype chettan and merlin chechi of Division of crop protection, ICAR-CTCRI for their valuable timely help during the difficult situation of the work.

I will always remember with gratitude to the helping hand of Ambu chettan throughout the work. I take immense pleasure to express my thanks to Jithu, Rahul, Amal, Ancy, Tom chettan, Vishnu chettan, Sreenath chettan and Abhishek for their support and advice during my research work.

My acknowledgement would be lacking if I don't mention my gratitude to my beloved friends Hasu, Paru, Raju, Adi, Jyoo, Ambu, Linu, Keeru, Musii, Vishnu, Alif, Anju, Neethu, Ami, Sree, Akku, Lachu for their invaluable care, constant support, motivation and selfless help.

I wish to express my deep gratitude to all the scientists and staff members of ICAR-CTCRI, teachers in college, my seniors and juniors for their timely support.

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

I owe this achievement to my beloved Achan, Amma, Chechi, Ettan, Ammutty 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.

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

Reshma T.K.

DEDICATED TO MY PARENTS

CONTENTS

LISTOF TABLES

LIST OF FIGURES

LISTOF PLATES

LIST OF ABBREVATIONS

%	Percentage
A_{260}	Absorbance at 260 nm wavelength
A_{280}	Absorbance at 280 nm wavelength
AFLP	Amplified fragment length polymorphism
bp	Base pair
CIM	Composite Interval Mapping
cM	CentiMorgan
CTAB	Cetyl Trimethyl Ammonium Bromide
DNA	Deoxy ribonucleic acid
DEG	Differentially Expressed Genes
EST	Expressed Sequence Tag
et al.	et alia
e-PCR	electronic PCR
F	Forward primer
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GGPS	Geranyl Geranyl Pyrophosphate Synthase
ICAR-CTCRI	Indian Council of Agricultural Research-Central Tuber Crops Research Institute
ISSR	Inter simple sequence repeat
KAU	Kerala Agricultural University
KEGG	Kyoto Encyclopedia of Genes and Genomes
NGS	Next Generation Sequencing
OFSP	Orange Fleshed Sweet Potato
PDB	Protein Data Bank

PCR	Polymerase chain reaction
QTL	Quantitative Trait Locus
RAPD	Random amplified polymorphic DNA
R	Reverse Primer
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RNase	Ribonuclease
RT-PCR	Real Time-polymerase chain reaction
sp.	Species
spp.	Species (plural)
SSR	Simple sequence repeat
Tm	Melting temperature
TrEMBL	Translated EMBL
USDA	United States Department of Agriculture
VAD	Vitamin A Deficiency
ZEP	Zeaxanthin Epoxidase

INTRODUCTION

1. INTRODUCTION

Root and tuber crops play a significant role in agriculture and facilitate food security in many developing countries. *Ipomoea batatas* (L.) Lam, commonly known as sweet potato belonging to the Convolvulaceae family, is a natural hexaploid root crop with chromosome number of 2n = 6X = 90. Sweet potato is an important food crop ranking seventh globally with 112.8 million tons (in 115 countries) production in 2017 and China is the leading producer, followed by Nigeria, Tanzania, Indonesia and Uganda (FAOSTAT, 2019). International Potato Centre reported that sweet potato is the third vital food crop in seven Central and East African countries. Sweet potato is praised as a "poor man's" crop as it characteristically grown and consumed by meager communities especially women- headed families. Sweet potato is considered as the food security crop due to its low agriculture input requirements and high yields in wider climatic conditions.

Sweet potato provides an excellent source of carbohydrates, dietary fiber, vitamins, micronutrients, antioxidants, such as phenolic acids, carotenoids, flavonoids, tocopherol and is low in fat and cholesterol. Different varieties of sweet potatoes are grown worldwide and these are generally characterized by the different flesh colours with varying phytochemical compositions.The flesh colour of the storage root varies from various shades of white, cream, yellow to dark-orange and purple depending upon the carotenoid and anthocyanin content. The predominant pigments and functional phenolic acids in sweet potato are the naturally strong free-radical scavengers, which attributes to many physiological functions including anti-oxidation, anti-tumor capacities, and prevention of cardiovascular diseases and vitamin A deficiency (Bovell Benjamin, 2007).

Orange Flesh Sweet Potato (OFSP) is appreciated due to its Vitamin A contribution and role in Vitamin A Deficiency (VAD) eradication in developing countries (Girard *et al.*, 2017; Kurabachew, 2015; Van Jaarsveld *et al.*, 2005).

Efforts are being taken by the International Potato Center (CIP) to replace the dominant white-fleshed varieties lacking β-carotene in Sub Saharan Africa (SSA) with provitamin A-rich OFSP. This have been considered as a breakthrough achievement as vitamin A deficiency affects more than 40% of children under the age of five in Sub Saharan Africa and is a leading cause of blindness and premature death (Mwanga *et al.*, 2011). The World Food Prize, 2016 awarded to four scientists who pioneered biofortification with OFSP highlights the importance of these efforts and the significance of OFSP in shifting human health outcomes.

The improvement of sweet potato varieties worldwide is facing major constraints owing to the lack of knowledge of the genetic, molecular, and physiological basis of key agronomic traits of this critical food crop (Wu *et al.*, 2018). Therefore, genomic data sources for sweet potato are greatly required for gene discovery and functional studies. Generating additional genomic resources would help in identifying the molecular basis of phenotypic variation and improve the design of efficient and effective marker-assisted breeding strategies (Yoon *et al.*, 2015).

Several studies concluded that inheritance of sweet potato flesh colour is a quantitative character controlled by several genes. Despite the significance of sweet potato as a source of β-carotene, relatively little research has focused on molecular aspects of carotenoid genes in this crop. The classical mass selection approach has been used for most of the sweet potato breeding programmes for high β-carotene cultivars which is slow and resource consuming. The candidate gene approach is a valuable method to investigate allelic variation involved in sweet potato storage root flesh pigmentation. A precise understanding of carotenoid biosynthesis and carotenogenic genes could contribute to metabolic engineering of sweet potato for use in sustainable development to solve the global food, energy, health and environmental problems (Kang *et al.*, 2017).

In non-model plants, it is difficult to identify the candidate genes involved in complex biosynthetic pathways due to the limited availability of genomic data.

High-throughput transcriptome sequencing and differential gene expression profiling are efficient and economic choices to characterize non-model organisms without a reference genome. In sweet potato, numerous genes were identified with roles in regulating the biosynthesis of anthocyanin's, starch, storage root formation and flower development, through transcriptome analysis (Tao *et al.*, 2013; Firon *et al.*, 2013; Xie *et al.*, 2012).

Quantitative trait loci (QTL) mapping is an efficient tool to uncover the genetic control of complex traits (Tecle *et al.*, 2010). QTL is defined as "a region of the genome that is associated with an effect of a quantitative trait. The major intention of QTL identification to locate useful genomic regions for use in marker-assisted selection (MAS) in breeding programmes and to identify the underlying genes responsible for desirable traits of interest. Integrating genetic (markers and QTLs) and genomic (gene and genome sequences) data deliver a complete catalogue of markers and putative candidate genes driving complex quantitative traits (Quraishi *et al.*, 2017).

The dissection of genetic architecture of sweet potato tuber colour into several chromosomal loci by QTL mapping and the combined use of the QTL mapping with transcriptome profiling represents a practical option to further refine the mapping resolution and identify potential candidate genes. The identification of genes and genomic regions that are involved in controlling agronomically important traits of sweet potato facilitates genetic studies for the development of marker-assisted breeding programs, thus helps to decipher the complex genetic structure of the crop. The present study was undertaken with the following objectives (i) to identify differentially expressed genes for various tuber colours in sweet potato (ii) to integrate QTL information on tuber colour with genomic information in sweet potato and (iii) to validate the identified candidate genes using accession of white and orange fleshed sweet potato.

REVIEW OF LITERATURE

21

2. REVIEW OF LITERATURE

Sweet potato (*Ipomoea batatas* (L.) Lam) is a clonally propagated crop native to the Americas. *Ipomoea* is the largest genus in the family Convolvulaceae, consisting of 600–700 species. Sweet potato is one of the widely cultivated species in the genus *Ipomoea* and consumed as a crop around the world. Sweet potato is a hexaploid species with a large genome size of 4.8–5.3 pg/2C nucleus and 90 chromosomes (2n = 6X = 90). Sweet potato is a herbaceous, perennial vine, bearing alternate heart-shaped or palmately lobed leaves and medium-sized sympetalous (fused petals) flowers. The edible tuberous root is long and tapered, with a smooth skin, whose colour ranges between red, purple, brown, and white. This crop is mainly used for human food (as such or in processed form), animal feed, and for manufacturing starch and its products. It is also an alternative source of bio-energy as a raw material for fuel production. This important root crop plays a critical role in food security, especially in developing countries.

2.1. SWEET POTATO PRODUCTION

Sweet potato (*Ipomoea batatas*) is one of the most important food crops in tropical and subtropical regions with an annual global production of approximately 100 million tons. Regardless of its origin in the tropical Americas, about 75% of sweet potato production now comes from Asian countries (FAO, 2015). Globally sweet potato is grown in 117 countries in an area of 8.62 million ha producing 105.19 million tons with a yield of 12.20 t ha -1 (FAO, 2016). Africa is the largest sweet potato cultivating region in the world with 95 percent of sweet potato production comes from developing countries, of which China having the majority share of 67.09 per cent (FAO, 2016). Sweet potato is an essential food source with very high production per capita across the relatively humid areas of Africa and provides more edible energy per hectare per day than wheat, rice, or cassava (FAO, 2014). In India, it is cultivated in almost all the states but major contribution comes from four states of Odisha, Kerala, West Bengal and Uttar Pradesh of which Odisha is the largest producer in India (FAO,

2016). Sweet potato is cultivated under an area of 0.13 million ha with a production of 1.47 million tonnes in India (FAO, 2016). The world production of sweet potatoes saw a gradual drop from 1993 to 2004, but has been stable for the past decade. The production reached over 100 million tonnes in 2014 (FAOSTAT, 2016).

2.2. NUTRITIONAL COMPOSITION OF SWEET POTATO

The sweet potato, because of its high nutritional composition and distinctive agronomic characteristics, has enormous potential to assist, deter and reduce food insecurity and mal-, under-, and over nutrition in developing and developed countries. Improved knowledge of the nutritional quality, utilization, and future economic significance of the crop has important implications for human food systems, nationally and internationally (Scott *et al.*, 2000). The sweet potato contains numerous nutrients including protein, carbohydrates, minerals (calcium, iron, and potassium), carotenoids, dietary fiber, vitamins (particularly C,A, folate, B6 and tocopherol), very low fat, and sodium. In addition to the nutritional values of sweet potatoes, it has been recognised as a functional food containing high concentration of multiple phytochemicals which might contribute to various health beneficial effects. Storage roots of sweet potato contains high levels of antioxidants such as ascorbate and carotenoids, is one amongst the healthiest foods, as well as one of the most effective starch crops for growth on marginal lands. Most studies on phytochemicals in roots and leaves of sweet potato indicates that the high level of polyphenols is associated with their health promoting and disease prevention especially, cancer-preventive effects of polyphenols in sweet potato have been widely investigated. The non-profit center for science in the public interest recently designated sweet potato as one of ten "super foods" because of the high levels of antioxidants, potassium and fibre, that can improve human health. The USDA reported that the sweet potato can generate two to three fold of carbohydrate as that of field corn, reaching the amount that sugarcane can produce, in Maryland and Alabama (Ziska *et al.*, 2009). The sweet potato leaves are consumed in Africa and Japan with protein content on a dry

weight basis has been reported to be as high as 27% (Diop, 1998). Tewe *et al.* (2003) reported that the protein content of sweet potato leaves as 18.4% and fiber content was between 3.3% and 6.0%. A sweet potato variety with high total polyphenol content and radical-scavenging properties than that of spinach, broccoli, cabbage, and lettuce has been reported (Ishiguro *et al.*, 2004).Islam (2006) reported that sweet potato leaves contain atleast 15 biologically active anthocyanin, that are beneficial to human health and may also be useful as natural food colorants. The major storage proteins in sweet potato sporamin A and B, which account for more than 80% of the total protein, are of great importance as they are proteinase inhibitors, having some anticarcinogenic properties has been reported (Maeshima *et al.*, 1985).The sweet potato tubers contain mostly substances of carbohydrate origin (sugars and fibers). Sweet potato leaves and shoots are excellent sources of vitamins A, C and B2 (riboflavin), and lutein according to Food and Agricultural Organisation (FAO, 2014).

2.3. COMPLEXITY OF SWEET POTATO CROP

Austin (1988) postulated that sweet potato originated in the region between Yucatan Peninsula of Mexico and the Orinoco River in Venezuela based on the analysis of key morphological characters of sweet potato and the wild *Ipomoea* species. The highest diversity of sweet potato has been found in Central America using molecular markers provides evidence that Central America is the primary centre of diversity and most likely the centre of origin, considering the richness of the wild relatives of sweet potato (Zhang *et al.*, 2000). However, no definitive conclusions are there regarding the evolutionary origin and genome structure of sweet potato. The recently proposed origin of sweet potato is of autopolyploid origin with *I. trifida* as the sole relative (Munoz *et al.*, 2018). Another hypothesis has proposed that *I.batatas* is an alloautohexaploid (2n = 6x = 90), resulting from an initial crossing between a tetraploid ancestor and a diploid progenitor followed by a whole genome duplication event with a B 1 B 1 B 2 B 2 B 2 B 2 genome composition (Magoon *et al.*, 1970; Yang *et al.*, 2016). It has also been suggested that hybridization by unreduced gametes of diploid *I. trifida* and a tetraploid *I.*

batatas is responsible for sweet potato origin (Freyre *et al.*, 1991). Though several species have been nominated as candidate ancestor species, a number of cytology and molecular studies have suggested that *I. trifida* is the closest wild species to sweet potato (Roullier *et al.*, 2013). In many non-model plant species, high-throughput and low-cost genome sequencing technologies have enabled the determination of whole genome sequences. The number of plant species for which the whole genomes have been sequenced is increasing annually. The sequencing of *I. batatas* genome as well as mapping studies for sweet potato is extremely cumbersome due to the extremely large genome and high heterozygosity. Accordingly, the genetic and genomic features of *I. trifida* have been studied as a potential reference for sweet potato.

The nature of the sweet potato crop is a hindrance to its genetic improvement. Some of the challenges of sweet potato breeding are its polyploidy nature, high heterozygosity, self and cross-incompatibility and large chromosome numbers (Cervantes-Flores, 2008; Chang *et al.*, 2009; Gurmu *et al.*, 2014). Mcharo and LaBonte (2007) observed that the inheritance studies in sweet potato are very complicated due to its hexaploid, self-incompatibility and heterozygous nature. These genetic conditions greatly affect the breeding and selection of sweet potato for quantitative traits of the crop, which slows down progress in genetic advances that can be attained within sweet potato breeding programmes (Mcharo and LaBonte, 2007). Moreover, Cervantes-Flores *et al.* (2011) indicated that the economically important traits in sweet potatoes follows quantitative inheritance due to the polyploid nature of the crop, and this makes their improvement difficult. Gasura *et al.* (2008) also commented that self- and cross-incompatibilities in sweet potato have remained major challenges in its breeding, and that these incompatibilities hinder production of segregating populations from specific crosses. However, attempts have been made to improve the crop production for different traits through estimation of combining abilities and thereby the type of gene action controlling the traits of interest, and the inheritance and heritability of distinct traits.

2.4. TUBER COLOUR VARIATIONS IN SWEET POTATO

Globally, many sweet potato cultivars exist exhibiting differences in size, skin colour (e.g., white, cream, yellow, orange, pink and red) and flesh colour (e.g., white, cream, orange, yellow, and purple) (Rose and Vasanthakaalam, 2011). Consumer acceptance mostly concerns the taste as well as the external appearance of the tubers, both of which are primarily influenced by their biochemical composition. Sweet potatoes are good sources of dietary fibre, minerals, vitamins, and antioxidants, such as phenolic acids, anthocyanins, tocopherol and β-carotene (Woolfe, 1992). Besides the role of antioxidants, carotenoids and phenolic compounds are responsible for the distinctive flesh colours in sweet potato. Various sweet potato cultivars having wide range of colours including white, yellow, orange and purple for skin and flesh of storage roots of sweet potato has been reported (Bovell Benjamin, 2007). These variations in the natural colours are mainly determined by the relative quantities of pigments carotenoids and anthocyanins. There are three major categories of sweet potato cultivars according to the variation in their pigments. The staple types are white, red/purple skinned with white/cream fleshed which are characterized by their high starch content. There are also the desert types which are orange skinned and orange fleshed with high β carotene content and the one with purple flesh colour. Different cultivars of sweet potato are also characterized by their colours, width, thickness and shapes of the leaves. Flavonoids, terpenoids, tannins, saponins, glycosides, alkaloids, steroids and phenolic acids are the major phytochemicals generally present in sweet potato. These constituents may affect the flesh and skin colours of the varieties. Orange varieties are especially rich in β-carotene, while purple sweet potatoes contain high content of anthocyanin than other sweet potato varieties. β-carotene is a terpenoid that are abundant in plants and fruits with a strongly coloured red-orange pigment. Anthocyanins belonging to the flavonoid group of phytochemicals are responsible for the purple and blue pigments in many fruit and vegetables. The antioxidant properties have mostly been attributed due to the anthocyanin and β-carotene contents of sweet potato. Phenolic acids such as chlorogenic, isochlorogenic, caffeic, cinammic, and hydroxycinammic acids are

also generally present in sweet potato. Phenolic acids have been associated with colour, sensory qualities, nutritional value as well as antioxidant activities. The phenolic acids are mostly present in purple fleshed sweet potato than in other colour varieties. Sweet potato varieties with orange and yellow flesh have been developed developed in Japan (Takahata, 2014). The cultivars that have been released are 'Sunny-Red'(Yamakawa *et al.*,1999), 'J-Red' (Yamakawa *et al.*,1998),'Hamakomachi' (Yoshinaga *et al.*, 2006) and 'Ayakomachi' (Kai *et al.*, 2004) with orange flesh and 'Tamaotome' (Ishiguro *et al.*, 2004), 'Benimasari' (Ishiguro *et al.*, 2004), 'Beniharuka'(Kai *et al.*, 2010), and 'Aikomachi' (Takada-Ohara *et al.*, 2016) with yellow flesh.

2.5. CAROTENOIDS IN SWEET POTATO

Four carotenoids are found in sweet potato leaves: lutein (47.6% of total carotenoids), β-carotene (25.2%), violaxanthin (13.9%) and neoxanthin (9.6%) (Chen and Chen, 1993). The composition of carotenoids in sweet potato leaves is very much similar to that of other plant chloroplasts (Botella-Pavia and Rodriguez-Concepcion 2006). The accumulation of natural pigments occurs in the storage root of sweet potato. The yellow or orange flesh colour intensity is directly associated with the carotenoid contents (Takahata *et al.*, 1993). The storage root is an excellent source of carotenoids because of the abundant concentration of major carotenoid trans-β-carotene, with the highest pro-vitaminA activity among the carotenoids (Bovell Benjamin, 2007). β-carotene, β-crypthoxanthin, zeaxanthin, violaxanthin and other unknown carotenoids are the major carotenoids found in the storage roots of sweet potato (Ishiguro *et al.*, 2010, Kim *et al.*, 2012, 2013, 2014). Many trans- and cis-forms of β-carotene, a major component in yellow- and orange-fleshed sweet potato, including β-carotene itself, 9Z-β-carotene, 13Z-β-carotene, β-carotene 5,8-epoxide, β-carotene 5,8;5′,8′-diepoxide (cis-isomer) and β-carotene5,8;5′,8′-diepoxide (diastereomer) has been reported. Several new carotenoids, ipomoeaxanthins A, B, C1 and C4 isolated from yellow-fleshed sweet potato were reported by Maoka *et al.* (2007). The β-carotene (80–92%) is the primary carotenoid in orange-fleshed sweet potato

and other carotenoids constitute less than 2% of the total in three different cultivars (Ishiguro *et al.*, 2010). The carotenoids showed anti-oxidative activities (Ishiguro *et al.*, 2010 & Oki *et al.*, 2006). Antioxidant activity of carotenoids have also been reported with preventive effects for some diseases in vitro and in animal models (Paiva and Russel, 1999). Epidemiological studies have shown that dietary carotenoids lead to lower risks for lifestyle- related diseases (Sugiura, 2015). Development of varieties with deep-orange or yellow flesh sweet potato is the prime objective of sweet potato breeders as the deep orange or yellow is an attractive colour in processed foods, and their carotenoids may contribute to the prevention of many diseases.

2.5.1. Carotenoid biosynthesis pathway

Carotenoids are a dietary source of provitamin A in both humans and animals. As the species are not capable to synthesize vitamin A, carotenoids are essential components of their diets. After ingestion, carotenoids are converted to rhodopsin and retinal, a visual pigment and precursor of retinoid acid, which controls growth, development and differentiation (Fraser and Bramley, 2004). Vitamin A deficiency causes night blindness, skin keratinization, dry eye syndrome, degenerative vision loss, impaired immune function and birth defects (Rao and Rao, 2007). Moreover, in biological membrane carotenoids directly scavenge reactive oxygen species. The carotenoid metabolic pathway and the function of the biosynthetic enzymes involved have been well studied in higher plants.

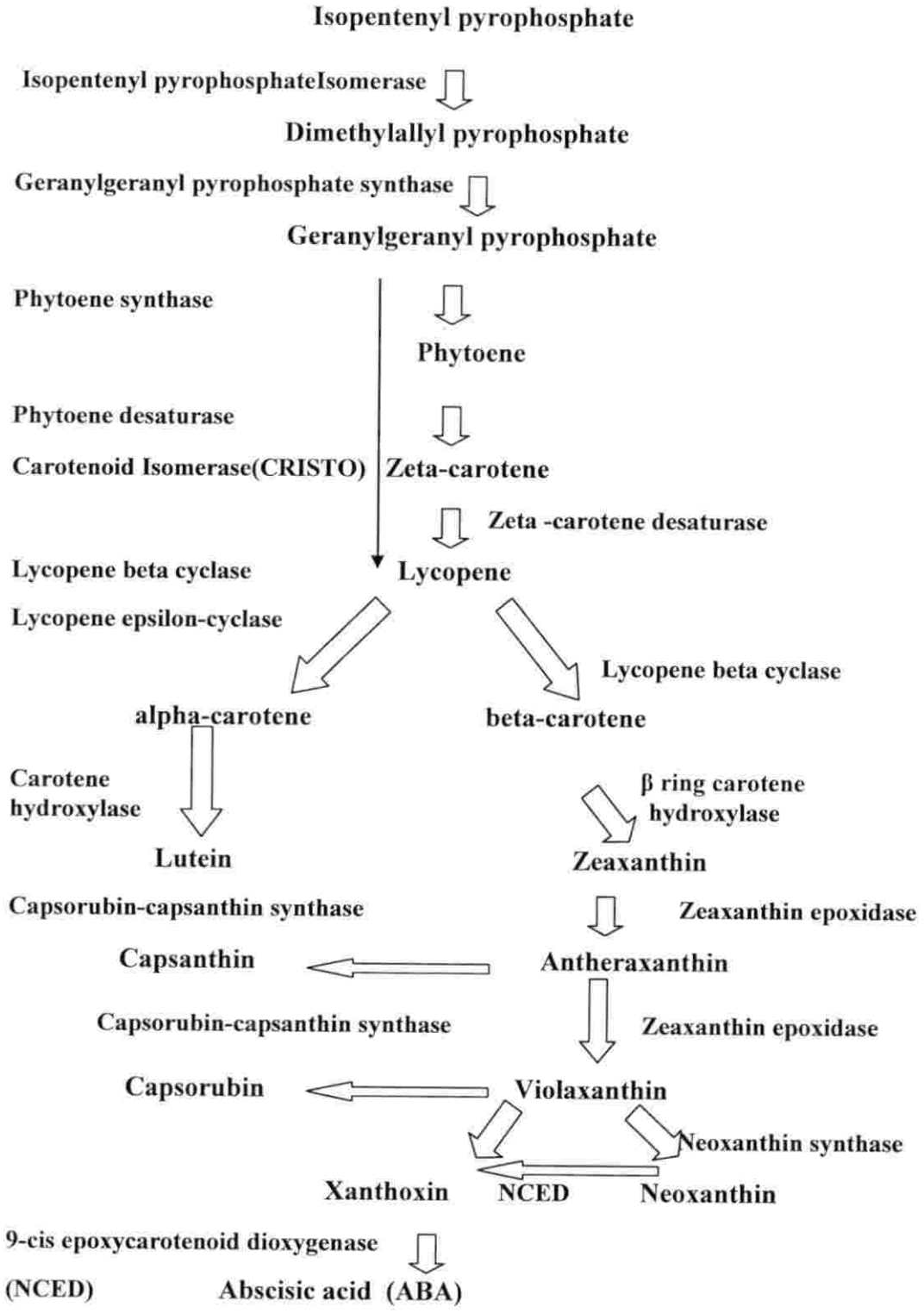


Figure 1. Carotenoid biosynthesis pathway (Cunningham and Gantt,1998)

2.5.2. Orange Fleshed Sweet Potato (OFSP)

Globally, sweet potato has a significant role in combating Vitamin A Deficiency (VAD). In developing countries, VAD is of public health concern, causing temporary and permanent eye impairments and increased mortality, particularly among children, pregnant, and lactating women. More than 230 million of the world's children have inadequate vitamin A intake, 13 million of whom are impacted by night blindness (Schweigert *et al.*, 2003). The habitual inadequate intake of bioavailable carotenoids (provitamin A) or vitamin A to meet physiological requirements is the main cause of VAD (Van Jaarsvld *et al.*, 2005). Plant foods do not contain vitamin A, however, they contain precursors or provitamin A (β-carotene and other carotenoids), that are converted into vitamin A by the human body. One strategy to control VAD is to improve dietary quality and quantity through diversification. Dietary diversification involves the development of β-carotene rich crops such as orange-fleshed sweet potato for human consumption (Van Jaarsveld *et al.*, 2005). In Kenya, orange-fleshed sweet potatoes have been recognized as the least expensive source of provitamin A throughout the year (Low, 1997). Another study reported that the consumption of orange-fleshed sweet potato could contribute significantly for reducing VAD in sub -Saharan Africa (Low *et al.*, 2017). The consumption of diets containing mainly orange fleshed sweet potatoes as a source of β-carotene considerably increased serum retinol (vitamin A status) levels of Indonesian children with marginal VAD (Jalal *et al.*, 1998). Van Jaarsveld *et al.* (2005) also concluded that enhanced orange fleshed sweet potato intake could be a feasible food-based approach for controlling VAD, particularly in children of developing countries. Carotenoids, with varying colours of red, yellow, or orange, are a diverse group of structurally related isoprenoids biosynthesized mainly by plants which have a capacity to trap lipid peroxyl radicals and singlet oxygen species (Ben-Amotz and Fishier, 1998). Provitamin A carotenoids are the ones that can be cleaved to produce retinaldehyde. β-Carotene, the main carotenoid in sweet potato, is

cleaved in the intestinal mucosa by carotene dioxygenase, yielding retinaldehyde, which is reduced to retinol (vitamin A). The total concentration of vitaminA in foods is expressed as microgram retinol equivalents. Nutritionally, 6 mg of dietary β-carotene is equivalent to 1 mg of retinol (Bender, 2002). The bioavailabity of provitamin A from orange-fleshed sweet potato appears to be more than that from dark green leafy vegetables (De Pee *et al.*, 1998). Several epidemiological studies have shown associations between carotenoids such as β -carotene, and reduced risk for cancer, heart diseases, and macula degeneration related with age (Kohlmeier and Hastings, 1995; Niizu and Rodriguez-Amaya, 2005). Infact the International Potato Center (IPC) attempts to substitute the β-carotene lacking white-fleshed varieties in SSA with provitamin A-rich orange-fleshed sweet potato (OFSP) have been regarded to be a breakthrough as vitamin A deficiency affects more than 40% of children under 5 years of age in SSA which is a major cause of blindness and premature death.

2.6. QUALITY IMPROVEMENT OF SWEET POTATO

Sweet potato is a vegetatively propagated auto-hexaploid, highly heterozygous, generally self- incompatible and outcrossing crop. Although many genes are available from other plant species, discovering new genes from sweet potato and its wild relatives may be useful for incorporating them for abiotic and biotic stresses resistance and high quality. The discovery of many of the important genes involved in controlling quantitative traits have been achieved through genomic approaches. The sweet potato genome is still unavailable. One of the efficient way for discovering and characterizing novel enzymes and transcription factors from sweet potato is sequencing of its transcriptome, which provides an important transcriptional data source for studying storage root formation, flower development, abiotic and biotic resistance, and starch, carotenoids, and anthocyanins biosynthesis and characterizing the important responsible genes of various traits of this crop (Firon *et al.*, 2013; Li *et al.*, 2015; Schafleitner *et al.*, 2010; Tao *et al.*, 2013; Wang *et al.*, 2010; Xie *et al.*, 2012). QTL analysis has become an efficient method for identifying responsible genes for different traits

by producing transformants of the corresponding DNA region. For further isolation and utilization, fine mapping of significant QTLs for storage root yield, nutritional quality and other characters is vital. The identification of agronomically important genes can be utilized for improvement of sweet potato by the introduction of the genes to commercial sweet potato cultivars.

2.7. TRANSCRIPTOME ANALYSIS OF SWEET POTATO

Transcriptome sequencing is one of the most efficient tools for gene discovery. The identification of putative genes and differential gene expression analysis has been made possible by large-scale *de novo* transcript sequencing. Due to the limited availability of genomic data, it is difficult to identify the candidate genes involved in complex biosynthetic pathways in non-model plants. This limitation has been overcome with high-throughput transcriptome sequencing technology, as it can generate large amounts of data on genome wide transcription. Several sweet potato transcriptomes have been sequenced, which act as a significant data source for storage root formation, flower development, and anthocyanin biosynthesis of the crop.

De novo transcriptome sequencing was done by using RNA extracted from tuberous roots of Jingshu 6, a purple sweet potato variety (Xie *et al.*, 2012). From the transcriptome analysis, 58,800 unigenes were obtained, among them 40,280 were identified as protein-coding genes. At least 3,553 genes were considered to be involved in the biosynthesis pathways of starch, alkaloids, anthocyanin pigments, and vitamins based on GO and KEGG analysis. The first digital gene expression (DGE) analysis in sweet potato was reported by Tao *et al.* (2012) by using Illumina GAII platform for seven tissue samples of young leaves, mature leaves, stems, fibrous roots, initial tuberous roots, expanding tuberous roots and harvest tuberous roots sequencing. Annotation of 128,052 transcripts were done by using OmicsBox , BLASTX, GO and KEGG analysis to identify the differentially expressed genes. A total of 58,800 unigenes were identified in a high-throughput RNA sequencing performed for comprehensively analyzing the transcriptome of the purple sweet potato. Majority of the unigenes identified were

protein-coding genes, with at least 3,553 genes participating in many important biological and metabolic pathways, including pigment biosynthesis. The key enzymes for synthesizing substrates of anthocyanin pigments in the pathway of flavonoid biosynthesis were discovered as chalcone isomerase genes and F3'H (Xie *et al.*, 2012). Transcriptome analysis of sweet potato cultivar, Xushu 18 flowers were done to identify the putative floral-specific and flowering regulatory-related genes by using the RNA-sequencing technique. A total of 2595 putative floral-specific and 2928 putative vegetative-specific transcripts were detected. A large number of transcripts similar to the key genes in the flowering regulation network of *Arabidopsis thaliana* were also identified (Tao *et al.*, 2013). Phenylpropanoid biosynthesis and its delivery into carbohydrate metabolism and starch biosynthesis are the major events involved in storage root initiation. In a transcriptome analysis of sweet potato root, specific genes in the phenylpropanoid pathways were pointed out, providing potential targets for sweet potato genetic engineering. The storage roots of orange-fleshed sweet potato typically have a high carotenoid content (Firon *et al.*, 2013). The transcriptome analysis of orange-fleshed sweet potato cultivar "Weiduoli" and its high carotenoid mutant "HVB-3" were sequenced. A total of 35,909 unigenes were harvested from Weiduoli and HVB-3. There were 874 DEGs between HVB-3 and Weiduoli, 401 of which were upregulated and 473 were downregulated in HVB-3 compared to Weiduoli. 22 DEGs related to carotenoid biosynthesis existed between Weiduoli and HVB-3. GGPS, GGPR, and DHDDS involved in terpenoid backbone biosynthesis were identified. GGPS is the key enzyme of carotenoid biosynthesis. GGPR converts geranylgeranyl diphosphate (GGPP), the precursor for carotenoid biosynthesis, to phytyl diphosphate in the tocopherol and chlorophyll biosynthetic pathways. DHDDS is involved in the biosynthesis of isoprenoids, which are the precursors of carotenoid biosynthesis (Li *et al.*, 2015).

2.8. GENETIC LINKAGE MAPPING AND QTL ANALYSIS

Molecular breeding, also called marker- assisted selection (MAS), refers to the method of using DNA markers which are tightly linked to particular traits to

assist phenotypic selection. Molecular breeding has several advantages over traditional breeding such as selection at seedling stage, no influence of environment, and selection of preferred homozygotes, thus improving the genetic improvement. It is now very easy to detect and characterize a large number of DNA markers with the rapid development of next-generation sequencing (NGS) technologies. The powerful tools to speed up genetic improvement of economically important traits through MAS are genomic resources such as molecular markers, linkage maps, ESTs and genome sequences, as well as mapped quantitative trait loci (QTL) for important traits (Ye *et al.*, 2017).

A genetic linkage map is an essential tool in molecular breeding for genetic improvement (Guimaraes, 2007). Genetic maps will facilitate genome mapping, genetic dissection of QTL and positional cloning of significant genes and provide a scaffold to assemble physical maps and an essential tool for functional genomics (Harushima *et al.*, 1998; Meksem and Kahl, 2006). For gene cloning, quantitative trait locus (QTL) detection, comparative genomic research, and marker-assisted breeding, construction of linkage map is essential. However, in sweet potato, linkage map construction has been challenging, due to the species complex genomic architecture: high heterozygosity, general self-incompatibility, and hexaploidy, with a large number of small chromosomes ($2n = 6x = 90$). Although genetic and linkage analysis in sweet potato has lagged far behind that in other plant species, both linkage map construction and QTL mapping in the species have been reported despite of its complex structure.

To date, several studies have reported construction of genetic linkage maps in sweet potato (Cervantes-Flores *et al.*, 2008, Chang *et al.*, 2009; Kriegner *et al.*, 2003; Li *et al.*, 2010; Monden *et al.*, 2015; Ukoskit and Thompson 1997; Zhao *et al.*, 2013). The first one was reported by Ukoskit and Thompson (1997), who constructed low-density linkage maps using randomly amplified polymorphic DNA (RAPD) markers, with a mapping population of progenies derived from 'Vardaman', exhibiting high early root yield, drought tolerance, and susceptibility to root-knot nematode (RKN), and 'Regal', which is resistant to disease, RKN,

and insects. Kriegner *et al.* (2003) developed linkage maps in an F 1 mapping population using AFLPs that was derived from a cross between 'Tanzania' and 'Bikilamliya' and obtained a total of 90 and 80 linkage groups, including 632 and 435 markers in the 'Tanzania' and 'Bikilamliya' maps, respectively, with a total lengths of 3655.6 and 3011.5 cM. Another set of linkage maps was developed by Cervantes-Flores *et al.* (2008) from a mapping population that was derived from a cross between 'Tanzania' and 'Beauregard' using AFLPs. In the 'Tanzania' map, 86 linkage groups, including 947 markers, were obtained with an entire length of 5792 cM and an average between-marker distance of 4.5 cM. Zhao *et al.* (2013) constructed the genetic linkage map in sweet potato with the highest marker density, using AFLP and SSR markers in an F1 mapping population of two cultivars 'Xushu 18' and 'Xu 781'. In the 'Xushu 18' map, 90 linkage groups were obtained, which included 2077 markers, and the total length of the maps was 8184.5 cM, with an average between-marker distance of 3.9 cM. Monden *et al.* (2015) detected comprehensive retrotransposon insertion polymorphisms in a NGS-based study in sweet potato, and developed genetic linkage maps.

QTL analysis has been carried out to detect the genetic bases of significant agronomic or physiological traits, offering valuable information for trait improvement. Genetic markers enabled the detection of QTLs that are significantly associated with traits thus increasing selection efficiency (Wang *et al.,* 2011). Quantitative trait loci (QTLs) are genomic regions (loci) associated to the phenotypic variation of a trait. In crop plants, quantitative variation is a feature of many agronomically important traits, such as yield, quality or disease resistance. Many economically important traits are quantitatively inherited, influenced by the environment and controlled by many genes of small and large effect. A powerful strategy to explore the genetic basis of such complex quantitative traits is through identification of quantitative trait loci (QTL) associated with the trait (Tecle *et al.*, 2010).

Several QTL mapping analyses have been conducted in sweet potato using the constructed linkage maps (Cervantes-Flores *et al.*, 2008; Zhao *et al.*, 2013) for

identifying agronomically important genes or genomic regions, and so far, such analyses have targeted economically important traits, such as nematode resistance, starch content, dry-matter content, β-carotene content, and root yield. The first QTL analysis in sweet potato was conducted by to identify target genes or chromosomal regions associated with RKN resistance, by using a mapping population of 'Tanzania' and 'Beauregard' by Cervantes-Flores *et al.* (2008). Cervantes-Flores *et al.* (2011) conducted QTL mapping analysis for dry-matter, starch, and β-carotene content using a mapping population of 240 plants derived from a cross between two parental cultivars 'Tanzania' and 'Beauregard' with significantly different traits, where the former is a white cream-colored African landrace that has a high dry-matter content (>30%), and the latter is an American orange-fleshed cultivar with a low dry-matter content. In the research, 13, 12, and eight QTLs were identified for dry-matter, starch, and β-carotene content, respectively. Zhao *et al.* (2013) developed the first map including 90 complete sweet potato linkage groups, and further analyses conducted by Yu *et al.* (2014) and Li *et al.* (2014) identified QTLs and co-localizing markers for agronomically important traits, such as dry-matter (Zhao *et al.*, 2013), starch content (Yu *et al.*, 2014), and yield (Li *et al.*, 2014). The analysis identified twenty-seven, eight, and nine QTLs for dry-matter content, starch content, and root yield respectively.

2.9. CANDIDATE GENE IDENTIFICATION

The anchoring of genetic maps and QTLs on reference genome sequence enables the identification of large genome regions containing hundreds of genes. One of the strategies to reduce this huge number of genes to a reasonable number of candidate genes, is to select candidate genes based on functional knowledge (for ex. known biosynthetic pathways; transcriptomic data; annotation inferred from homologous genes in other species) and check if the candidates co-locate with QTLs (Street *et al.*, 2006). There has also been an increasingly desire to identify the candidate genes underlying the QTLs responsible for traits of interest. The most common method to refine candidates underlying a QTL is to search for physically-proximate genes with annotations or gene ontology reflecting the trait

of interest. Another strategy, is to combine QTL position, structural information and transcriptomic experiments to define a gene list for functional characterization (Ranjan *et al.*, 2010). Today, the biological interpretation of candidate gene lists is made possible by the accessibility of biological knowledge accumulated in public databases (e.g. Gene Ontology) and high- throughput bioinformatic enrichment tools. The development of enhanced crop varieties through traditional breeding can be improved by understanding the response of QTL in different environments or genetic backgrounds. If the genes underlying the QTL are known, then it is also possible to use transgenic approaches for directly introducing beneficial alleles across wide species boundaries.

Quantitative trait loci (QTL) analysis is a promising initial step in the identification of genes responsible for carotenoid accumulation in plants (Liu *et al.*, 2003; Pflieger *et al.*, 2001). The position of QTL controlling carotenoid content and flesh colour in corn (Wong *et al.*, 2004) and carrot (Just *et al.*, 2007) have been compared to identify markers of the candidate genes that have potential for marker-assisted selection (MAS). In poplar species, 77 QTLs controlling 11 traits were identified, where 58 QTLs confidence intervals among them could be projected on the reference genome. Functional annotation was further done based on the data retrieved from the plant genome database Phytozome and from an inference of function using homology between *Populus* and the model plant *Arabidopsis*. Genes located within QTL confidence intervals were retrieved using enrichments in gene ontology (GO) terms (Monclus *et al.*, 2012).

2.9.1 Functional annotation of the candidate gene

Genome annotation is a dynamic process of gaining additional information on molecular and genome biology. Most genomic projects use DNA and/or RNA sequencing to carry out different kinds of functional studies. In these studies, reference genomic databases are useful for annotation of identified genomic elements. Nowadays, many integrated bioinformatic resources are available online for researchers to utilize genomic information in their own projects. The most popular genomic resources include Ensembl (Flicek *et al.*, 2013) and UCSC

(Karolchik *et al.*, 2014). Several sources of information including known EST, mRNAs, genes and protein information are used to annotate new genes in different species. The *ab initio* tools for gene identification, such as GenScan (Burge & Karlin, 1997) and N-Scan (Gross &Brent, 2006) are also used for newly sequenced species. These genome browsers deliver fundamental information for genomes including e.g. genes, genetic variation, regulatory elements and conservation. To functionally characterize genes and gene products, Gene Ontology database (Ashburner *et al.*, 2000) is widely used in life sciences. This database provides consistent biological ontology that can be widely applied to any species. The structured and controlled vocabulary of GO terms makes it possible to consistently describe the biological functions of gene products (Yu& Hinchcliffe, 2011). Biological pathways are involved in metabolism, gene regulation and signal transduction. These pathways are comprised of interacting biological molecules to perform different functions. Many of these pathways are curated in Kyoto Encyclopaedia of Genes and Genomes (Kanehisa *et al.*, 2012). The genes in selected genomic regions within QTLs were mapped to several KEGG pathways to reveal biological interaction networks that underlie interacting loci.There are many bioinformatic resources available for genomics research and investigating them individually for annotation of a large set of genes is laborious and very time consuming. The Database for Annotation, Visualization and Integrated Discovery (Huang *et al.*, 2008) provides functional analysis tools for batches of genes. A gene list can be submitted to DAVID for annotation enrichment analysis. OmicsBox is another suite which serve as a comprehensive bioinformatics tool for functional annotation of sequences and data mining on the resulting annotations, primarily based on the gene ontology (GO) vocabulary. The major OmicsBox strength is the combination of functional annotation and data mining on annotation results, which means that, within one tool, researchers can generate functional annotation and assess the functional meaning of their experimental results. Functional annotation of putative QTLs controlling black tea quality and percent relative water traits in two tea populations and mapping of gene ontology (GO) terms was done with the OmicsBox program (OmicsBox

v3.2). Functional annotations of detected QTL led to the identification of functions of 37 putative candidate genes in parental clones that could be involved in the expression of the targeted traits (Koech *et al.*, 2019). The integration of QTL mapping of functional traits, genome annotation and allele association yielded several candidate genes involved with molecular control of photosynthesis and water use efficiency in response to drought in *pinus pinaster* has been done with OmicsBox software (de Miguel *et al.*, 2014).

2.8. INTEGRATED TRANSRIPTOME ANALYSIS AND QTL MAPPING TO IDENTIFY CANDIDATE GENES

The combination of linkage mapping and transcriptome analysis has greatly improved the discovery of candidate gene underlying the QTL. The integrated analysis of linkage mapping and transcriptome profiling has emerged as a useful tool to reduce the number of candidate genes underlying various complex traits in a variety of plant species, such as *Arabidopsis* (Xu *et al.*, 2015), maize (Marino *et al.*, 2009), and rice (Deshmukh *et al.*, 2010; Yano *et al.*, 2012).

Candidate genes underlying QTLs associated with Low-P tolerance and response mechanisms to Low-P stress in soybean were identified using QTL mapping and transcriptome profiling. The information explaining the effects of Low-P on growth and development in soybean can be improved by characterization of identified genes and other QTL-colocalized DEGs and DEG-encoding phosphatases. Four candidate genes in QTL regions associated with salt tolerance in common wild rice were identified through integrating RNA sequencing and QTL mapping (Zhang *et al.*, 2017). In *B. napus,* a total of 31 QTLs for the leaf morphological traits were evaluated. In addition, 74, 1,166 and 1,272 DEGs were identified using RNA-Seq technologies and 1,205 genes were identified in QTL regions by aligning the SNP marker physical locations with the oilseed reference genome. Thirty-three candidate genes were identified by QTL mapping and RNA-Seq technologies serve as a valuable source for studies of the genetic control of the regulation of leaf morphological traits in *B. napus* (Jian *et*

al., 2019). In another study, QTL mapping and RNA sequencing were done to investigate the genetic and molecular mechanism of pod number variation in rapeseed. Fifteen consensus QTLs were identified through meta -analysis and GO and KEGG analysis of 9135 DEGs between the shoot apical meristem of two parent cultivars provided new insights into pod number variation (Yang *et al*., 2016). Another research in sorghum reported the DEG transcripts between parents and RIL bulks which overlaid to the identified QTL confidence intervals using physical positions to identify candidate DEG transcripts associated with QTLs for agronomic traits expressed under the N- stress (Gelli *et al*., 2017)

Genetic improvement of sweet potato through traditional plant breeding is difficult due to its polyploid nature, genetic complexity, and high variability with regard to flower production and incompatibility. Generating additional genomic resources would aid efforts to identify the molecular basis of phenotypic variation and advance the design of efficient and effective marker -assisted breeding strategies.

MATERIALS AND METHODS

3. MATERIALS AND METHODS

The study entitled "Integration of Quantitative Trait Loci (QTL) for tuber colour variations with genomic information in sweet potato (*Ipomoea batatas* L.)" was conducted at ICAR-CTCRI during October 2018 to August 2019. This chapter is presented with the detailed information about experimental materials and methodologies adopted for handling various experiments.

3.1. IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES FROM RNA SEQ DATA

The preliminary transcriptomics data of sweet potato was collected from ICAR-CTCRI Sreekariyam, Thiruvananthapuram. Total RNA from the sweet potato tubers were extracted and Paired End (PE) libraries were prepared from Tuber samples using TruSeq stranded mRNA Library Prep Kit. The libraries were sequenced on NextSeq 500 platform. A total of six tuber samples along with their replicates were used for the study. The tubers samples of two white fleshed, two orange fleshed and two purple fleshed varieties were chosen for the analysis. The details of the six tuber samples and their replicates are shown in Table 1.

Sample	Flesh colour
Tuber 1R1	White
Tuber 1R2	White
Tuber 2R1	White
Tuber 2R2	White
Tuber 3 R1	Purple
Tuber 3 R2	Purple
Tuber 4 R1	Purple
Tuber 4 R2	Purple
Tuber 5 R1	Orange
Tuber 5 R2	Orange
Tuber 6 R1	Orange
Tuber 6 R2	Orange

Table 1. Tuber samples taken for study

BIOINFORMATICS ANALYSIS

3.1.1. Raw sequence processing and quality control

FastQC

The present study the bioinformatics tool FastQC (Version 0.11.7) was used for the quality checks. FastQC provide a simple way to analyse quality control checks on raw sequence data from high throughput sequencing pipeline. The tool command lines are exemplified using a paired-end (PE) FASTQ formatted file set, such as read_1.fq.gz and read_2.fq.gz. The raw data received from the sequencing facility will be in these files. FastQC are check the overall sequence quality, the presence or absence of overrepresented sequences and the GC percentage distribution. The output of FastQC is a zip archive containing an

HTML document, which is sub-divided into sections describing the specific metrics that were analyzed.

i. **Basic statistics**

The Basic Statistics module generates some simple composition statistics for the raw sequence file analysed. Here, the file name, file type, total sequences, filtered sequences and percentage of GC content will be displayed.

ii. **Per base sequence quality**

This gives an overview of the range of quality values across all bases at each position in the FastQ file.

iii. **Per sequence quality scores**

The per sequence quality score report allows to identify if a subset of the sequences have universally low quality values.

iv. **Per base sequence content**

Per base sequence content gives the proportion of each base position in a file for which each of the four DNA bases has been called.

v. **Per base GC content**

Per Base GC Content plots out the GC content of each base position in a file.

vi. **Per sequence GC content**

This module measures the GC content across the whole length of each sequence in a file.

vii **Overrepresented kmer sequences**

Overrepresented sequences will spot an increase in any exactly duplicated sequences. A k-mer is a motif of length 'K' observed more than once in a sequenced sequence.

3.1.2. Quality trimming and adapter removal

TRIMMOMATIC

Trimmomatic is a fast command line tool that can be used to trim and crop FASTQ data as well as to remove adapters. Trimmomatic works with FASTQ files (using phred + 33 or phred + 64 quality scores). The Phred scale quality represents the probability that the base call is incorrect. A Phred score of 10 corresponds to one error in every 10 base calls or 90% accuracy; a Phred score of 20 to one error in every 100 base calls or 99% accuracy.phred33 or -phred64 specifies the base quality encoding. If no quality encoding is specified, it will be determined automatically. Files compressed using either "gzip" or "bzip2" are supported, and are identified by use of ".gz" or ".bz2" file extensions.

Command for single end data:

For single end data, one input and one output file are specified.

java-jar <path to trimmomatic jar> SE-phred33input.fq.gzoutput.gq.gz ILLUMINACLIP:/adaptors/TruSeq3-SE.fa: 2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW: 4:15 MINLEN:36

Command for paired end data:

For paired end data, we have to specify 2 input files which generate 4 output files (2 forpaired output and the other 2 corresponding to unpaired output data).

java-jar<path to trimmomatic.jar> PE phred33input1.fq input2.fq. Output1paired.fq.gz. Output1unpaired.fq.gz output2paired.fq.gz output2unpaired.fq.gz ILLUMINACLIP:/adaptors/TruSeq3- PEfa: 2:30:10 LEADING:3 TRAILING:3SLIDINGWINDOW: 4:15 MINLEN:36

i. ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
ii. SLIDINGWINDOW: Performs a sliding window trimming approach. It starts scanning at the 5' end and clips the read once the average quality within the window falls below a threshold.

iii. LEADING: Cut bases off the start of a read, if below a threshold quality
iv. TRAILING: Cut bases off the end of a read, if below a threshold quality
v. MINLEN: Drop the read if it is below a specified length
vi. TOPHRED33: Convert quality scores to Phred-33
vii. TOPHRED64: Convert quality scores to Phred-64

3.1.3. *De novo* transcriptome assembly

TRINITY

The software called trinity was used for the *de novo* trancsriptome assembly. *De novo* assembly is performed when there is no reference sequence available. Trinity represents a novel method for the efficient and robust de novo reconstruction of transcriptomes from RNA-seq data. Trinity combines three independent software modules: Inchworm, Chrysalis, and Butterfly, applies in an order to process large volumes of RNA-seq reads. Trinity arranges the short sequences into contiguous sequences representing gene transcripts.

i. Inchworm divides the reads based on commonly found sequence patterns and assembles initial contigs.
ii. Chrysalis, which groups related contigs (such as splice forms) together into connected de Bruijn graphs. In de Bruijn graph, short reads were broken down into short sequences of length *k*, referred as *k*-mers. *k*-mers are then used to form the graph.
iii. Butterfly, which collapses these graphs and aligns the reads to them, distinguishing between splice forms and paralogs.

Command line for trinity

Trinity –seqType fq –left output_forward_paired.fastq —right output_reverse_paired.fastq–CPU 8 –max_memory 100G

The assembled transcripts will be found at 'trinity_out_dir/Trinity.fasta'.The filteredhigh quality reads of all the six samples were assembled into transcripts using TrinityRNA -Seq assembler on default parameters.

3.1.4. Differential expression analysis

OMICSBOX

OmicsBox is an application for the functional annotation, management, and data mining of novel sequence data through the use of common controlled vocabulary schemas. The OmicsBox annotation procedure consists of three major steps: blast to find homologous sequences, mapping to collect GO terms associated to blast hits, and annotation to assign functional information to query sequences. The steps involved in the differential expression analysis:

3.1.4.1. Transcript level quantification

Transcript omics

The transcript-level quantification tool is designed for estimating gene and isoform expression levels from RNA-Seq data. It expects the sequencing reads in FASTQ format and it supports both single-end and paired-end data. A Count table is generated and it is used to perform a differential expression analysis within OmicsBox. The application is based on RSEM, a software package that quantifies expression from transcriptome data. The first step in the estimation of the expression levels of the Trinity reconstructed transcripts is the alignment of the original RNA seq reads back against the Trinity transcripts. Then RSEM is run to estimate the number of RNA seq fragments that map to each contig. Because the abundance of individual transcripts may significantly differ between samples, the reads from each sample must be examined separately, obtaining sample specific abundance values.

I. Run Create Count Table

This functionality can be found under **rna-seq → Create Count Table, Transcript-level Quantification** option.

Input Data parameters:

- **Input Reads:** The files containing sequencing reads. These files must be in FASTQ format or compressed FASTQ format (.gz). The raw sequenced data of all the six tuber samples and its replicates were taken as the input reads for creating the count table.
- **Transcript References:** The tool works with a set of transcripts sequences instead of a genome, such a file could be obtained from a reference genome database or a *de novo* transcriptome assembler. The FASTA files containing the sequences of reference transcripts of the tuber samples and replicates obtained from *de novo* assembly from Trinity was chosen as the transcript references.

Advanced Configuration parameters

- **Gene-level Estimations:** This option allows to estimate expression both at gene-level and isoform-level.
- **Transcript to Gene Map File:** Provide a file with information to map from transcript (isoform) identifiers to gene identifiers. Each line should be of the form: gene id transcript id, with the two columns separated by a tab character. The create transcript to gene map file, RSEM provides a perl script to generate transcript-to-gene map file from the fasta file produced by Trinity for each of the tuber samples

Usage:

Extract-transcript-to-gene-map-from-trinity trinity_fasta_file map_file

- Trinity_fasta_file: fasta file produces by trinity containing all transcripts assembled.
- Map_file: transcript-to-gene-map file's name.

3.1.4.2. Run differential expression analysis

OmicsBox tool is designed to perform differential expression analysis of count data arising from RNA-seq technology. This application, based on the edgeR program, allows identification of differentially expressed genomic features in a pairwise comparison of two different experimental conditions. EdgeR works on a table of integer read counts, with rows corresponding to genes and columns to independent libraries. edgeR is mainly concerned with relative changes in expression levels between conditions. The program normalizes for RNA composition by finding a set of scaling factors for the library sizes that minimize the log fold changes between the samples for most genes. The default method for computing the scale factors uses a trimmed mean of M-values (TMM) between each pair of samples. The different steps are as follows:

I. Load Data

Go to **File → Load → Load Count Table** and select the .txt file containing the count table in tab-delimited format.

II. Run Pairwise Differential Expression Analysis

rna-seq → Run Differential Expression Analysis and choose the "Pairwise Differential Analysis" Option.

Default parameters:

Preprocessing Data Page:

- Count-Per-Million (CPM) Filter: 0
- Samples reaching CPM Filter: 1
- Calculate normalization factors to scale the raw library sizes: Yes

Experimental Design Page

- **Experimental design file:** Select.txt file containing experimental factors with the experimental conditions associated to each sample in tab-delimited format. (Table 2).

Table 2. Experimental design file for pairwise differential expression analysis

Samples	Condition	Individual
Tuber4_R2	Purple	4
Tuber3_R1	Purple	3
Tuber4_R1	Purple	4
Tuber5_R2	Orange	5
Tuber2_R1	White	2
Tuber3_R2	Purple	3
Tuber2_R2	White	2
Tuber1_R1	White	1
Tuber1_R2	White	1
Tuber6_R1	Orange	6
Tuber6_R2	Orange	6
Tuber5_R1	Orange	5

Comparison and Test Page

- Design Type: Simple design type
- Statistical Test: GLM likelihood ratio

The pairwise analysis was done between all the three tuber (condition- colour) samples with differing the primary target.

Pairwise analysis between orange and white:

- Primary experimental factor: Condition
- Primary contrast condition: Orange

- Primary reference condition: White

Pairwise analysis between orange and purple:

- Primary experimental factor: Condition
- Primary contrast condition: Orange
- Primary reference condition: Purple

Pairwise analysis between purple and white:

- Primary experimental factor: Condition
- Primary contrast condition: Purple
- Primary reference condition: White

3.1.4.3. Enrichment Analysis

Functional enrichment analysis is a procedure to identify functions that are over-represented in a set of genes and may have an association with an experimental condition. Enrichment analysis perform a functional enrichment analysis from the pairwise differential expression project. The Functional Analysis Module offers two different statistical tests, the Fisher's Exact Test and Gene Set Enrichment Analysis.

I. Fisher's Exact Test

This functionality can be found under *Functional Analysis → Enrichment Analysis → Enrichment Analysis* or perform a functional enrichment analysis from the pairwise differential expression project in Omicsbox. The default parameters are:

Reference annotation: The annotated file of fasta file generated from trinity for the tuber samples.

- Filter value: 0.05
- Filter mode: FDR
- Annotations: GO

- GO categories: Biological process, Molecular function, Cellular component.

II. Gene Set Enrichment Analysis (GSEA)

OmicsBox includes the GSEA computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states. Genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric. This functionality can be found under *Functional Analysis → Enrichment Analysis → Gene Set Enrichment Analysis (GSEA)* or perform a functional enrichment analysis from the pairwise differential expression project. The default parameters used are:

- Number of permutations: 1000
- Enrichment statistic: Classic
- GO category: Biological process, Molecular function, Cellular component
- Filter mode: FDR
- Filter value: 0.25

3.1.4.4. Functional annotation analysis of the core enriched transcripts

I. Blast:

OmicsBox uses the Basic Local Alignment Search Tool (BLAST) to find sequences similar to the query set. BLAST can be performed in different actions:

i. CloudBlast: This is a feature for massive sequence alignment tasks. CloudBlast is a high-performance, secure and cost-optimized solution for the analysis. This service is totally independent from the NCBI servers and provides fast and reliable sequence alignments.

ii. QBlast@NCBI: NCBI offers a public service that allows searching molecular sequence databases with the BLAST algorithm.

iii. Local BLAST against own database: It is used to BLAST to query a local/own database

The parameters used in blast:

Blast Configuration Page parameters

- e-mail address in case of using the NCBI BLAST web service
- BLAST program: blastx-fast
- BLAST DB (Database): nr
- Taxonomy Filter: No filter
- BLAST expect value: 1.0E-3
- Number of BLAST hits: 20

Advanced Configuration Page

Blast parameters

- Word size: 6
- Low complexity filter
- Filter Options:
- HSP length cutoff: 33
- HSP-Hit Coverage: 0

The results of the BLAST queries can also be directly saved to a file in different formats of xml, txt, html.

II. Gene Ontology Mapping

Mapping is the process of retrieving GO terms associated to the Hits obtained by the BLAST search. BLAST result accessions are used to retrieve gene names or symbols making use of two mapping files provided by the NCBI. Identified gene names are then searched in the species specific entries of the gene-product table of the GO database. GeneBank identifiers (gi), the primary blast Hit ids, are used to

retrieve UniProt IDs making use of a mapping file from PIR (Non-redundant Reference Protein Database) including PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept and PDB.

III. Gene Ontology Annotation

This is the process of selecting GO terms from the GO pool obtained by the Mapping step and assigning them to the query sequences.

The default parameters for GO annotation.

- Annotation Cut-Off (threshold): 55
- GO-Weight:5
- E-Value-Hit-Filter: 1.0E-6
- Hsp-HitCoverage CutOff: 0
- Hit Filter: 500

3.2. INTEGRATION OF QTLs FOR TUBER COLOUR VARIATIONS WITH GENOMIC INFORMATION IN SWEET POTATO

The study aims to identify putative candidate functional genes by investigating tuber flesh colour QTLs regions in sweet potato using the markers flanking the QTL region by applying computational approach. Mapping positions of these candidate genes are expected to be useful for sweet potato breeders to improve sweet potato varieties with varying flesh colours.

3.2.1. QTLs Mining

To date, only limited studies have been reported on identification of QTLs controlling the tuber flesh colour in sweet potato. Chang *et al.* (2009) reported QTLs associated with yield traits in sweet potato. The yield traits studied are top weight, root weight, root number, root shape, root skin colour, and flesh colour in the sweet potato. In the study, four QTLs for sweet potato root tuber flesh from two mapping populations of 120 F1 plants were derived from a reciprocal cross between two varieties, 'Nancy Hall' (NH) and 'Tainung 27'(TN27). TN27 produces an extremely high quantity of pollen whereas NH, has a high yield of storage roots with high levels of β-carotene. Genetic linkage map was constructed for various traits using Inter Simple Sequence Repeat (ISSR) markers. Cervantes-Flores *et al.* (2011) identified QTLs for dry-matter, starch, and β-carotene content in a hexaploid sweet potato mapping population derived from a cross between Tanzania, a white-fleshed, high dry-matter African landrace, and Beauregard, an orange- fleshed, low dry-matter sweet potato cultivar popular in the USA. The mapping population consisted of 240 progenies using AFLP markers and eight QTLs were identified for β-carotene content. The data of the QTLs for β carotene are represented in Table 3.

Trait	Cultivar	QTL name	Linkage group	Closest marker	LOD	Phenotypic variance (R^2)	Reference
Root flesh colour	Nancy Hall	FC1	NH Group 1	ISSR 809-7	2.7	10.7	Chang *et al.* (2009)
		FC2	NH Group 2	ISSR 811-1	5.9	27.8	
		FC3	NH Group 6	ISSR 809-14	2.5	14.1	
		FC4	NH Group 7	ISSR 825-16	2.5	20.4	
	Tainung 27	FC1	TN27 Group 3	ISSR 810-4	3.9	29.9	
		FC2	TN27 Group 12	ISSR 857-6	3.8	22.8	
β-Carotene content	Beauregard	caro1	B04.23	E43M5403	-	-	Cervantes-Flores *et al.* (2011)
		caro2	B08.48	E38M3725	-	-	
		caro3	B11.62	E36M5103	-	-	
		caro4	B12.69	E44M4902	-	-	
	Tanzania	caro5	T13.74	E45M3611	-	-	
		caro6	T13.76	E40M3105	-	-	
		caro7	T78	E46M3901	-	-	
		caro8	T82	E36M4015	-	-	

Table 3. Summary of QTLs identified for root flesh colour in sweet potato

Development of genetic linkage map in sweet potato based on SSR markers and the identification of QTLs related to β carotene, dry matter and starch content was done at ICAR-CTCRI by Nair *et al.* (2017). A mapping population of 208 progeny derived from a cross between 'ST-14' (female) and 'S-1'(male) were used in the study. The 'S-1' male parent is a spreading type with green colour vine and emerging leaf. The tubers are elliptical in shape with purple skin and white flesh colour. The dry matter content of the tuber was 37 percent and very sweet in taste. The female parent ST-14, is a semi erect type and tubers are ovate type with cream colour skin and dark orange flesh colour with dry matter of 27 percent and contain high β carotene content.

SSR (Buteler *et al.*, 1999) and EST-SSR (Wang *et al.*, 2011) primers were used for genotyping. Single marker analysis (SMA) was carried out using the phenotypic scores and the marker segregation patterns of 208 progenies to identify the SSR markers linked to β- carotene. The SMA between marker and phenotype resulted in the identification of six SSR markers having association with the β- carotene. The QTL analysis was done by both Simple Interval Mapping (SIM) and Composite Interval Mapping (CIM). Interval mapping uses probability estimates for the genotypes in intervals between markers. The information about the identified QTLs in the study are shown in Tables 4 & 5.

Table 4. QTLs identified for β-carotene by simple interval mapping

QTL name	QTL position	LOD score	R^2	Flanking markers
QTL	156.0	5.13	0.00	IB1809c-IB242b
QTL2	197.3	10.00	10.00	GDS0215a-GDS1059a
QTL3	216.3	11.84	0.00	IBSSR04c-GDS1059a
QTL4	728.1	7.99	0.008	IB318b-IB318a
QTL5	777.1	10.95	0.00	IBSSR04b-GDS0997d
QTL6	115.0	3.42	0.004	IBSSR04e-IBSSR04d

Table 5. QTLs identified for β-carotene by composite interval mapping

QTL name	QTL position	LOD score	R^2	Flanking markers
QTL1	27.0	5.63	0.008	IB1809c-IB1809b
QTL2	58.0	3.53	0.001	IB1809b-IB242b
QTL3	220.9	8.12	0.875	GDS1059a-IB1809a
QTL4	540.8	4.20	0.003	GDS0215b- GDS0134
QTL5	726.1	5.37	0.030	IB318b-IB318a
QTL6	778.1	8.95	0.00	IBSSR04b-GDS0997d
QTL7	116.0	2.57	0.002	IBSSR04e-IBSSR04d

LOD- Likelihood of odds score

Phenotypic R^2: This value indicates the relative importance of QTL influencing a trait. It is the percent of a total phenotypic variance for the trait that is accounted for by a marker

3.2.2. Sequence retrieval

Even though the sweet potato genome has not been fully public, sequences provided by the NGS sequencing core facility, MPI Molecular Genetics, Berlin has been utilised for the alignment of marker sequences. This sweet potato genome browser contains the half haplotype-resolved hexaploid genome of sweet potato which represents the first successful attempt to investigate the complexity of chromosome sequence composition directly in a polyploid genome, using sequencing of the polyploid organism itself rather than any of its simplified proxy relatives. The other genomic resources available for sweet potato to facilitate genome-enabled breeding in hexaploid sweet potato is derived from Genomic Tools for Sweet potato Improvement project that have generated the sequence for two diploid, highly inbred *Ipomoea* species: *Ipomoea trifida* (NSP306) and *Ipomoea triloba* (NSP323) to serve as a reference sequence for the hexaploid sweet potato genome.

The alignment of the flanking marker sequences (EST-SSR) with the genome sequences that are publically available with sweet potato genome assembly available at Ipomoea genome hub database and sweet potato genomics resource database. The marker sequences (ISSR and AFLP) linked to root flesh colour in sweet potato reported by Chang *et al.* (2009) and Cervantes-Flores*et al.* (2011) is not currently available on the databases and other sources.

Among the SSR markers, used for the identification of QTL for β-carotene by Nair*et al.* (2017), only EST-SSR marker sequences were able to retrieve from the study on characterization and development of EST-SSR markers in sweet potato(Wang *et al.*, 2011).

The SSR markers linked to β-carotene other than EST-SSR markers whose sequences which are not available were retrieved from the sweet potato genomics resource database which contains the genome sequence of two diploid *Ipomoea* varieties, *Ipomoea trifida* and *Ipomoea tribola*. The forward and reverse primers

of these SSR primers were available. In the database, there is an option available called *Ipomoea* ePCR tool. The *Ipomoea* e-PCR tool uses the program ipcress to perform an in-silico PCR simulation (e-PCR). The program works by searching for primer alignments in a reference genome that produce a product within the parameters selected.

Parameters for e-PCR tool

i. Primer Sequence: min length 10 nts, max length 30 nts, primers should be selected
ii. Product Size Span: min 1 bp, max 1000 bp
iii. Max Mismatches per Primer Alignment: min 0, max 2
iv. The *I. trifida* and *I. triloba* v3 pseudomolecules are the only reference genomes available.

The details of the SSR marker sequences that are linked to the QTL for β-carotene are shown in Table 6.

Markers	Forward primer (5'-3')	Reverse primer (3'-5')	Product size (bp)	SSR containing sequence
GDS0215	TGGTGCCTCAGGTGGACAAC	GGAGGATGGAGGAGATCGGATG	135	TGGTGCCTCAGGTGGACAACGTCGTCGTTCGCGTTGACTGCGTGATCTCCGACGAAGACGGTTCTTCGCCGGCGCCGGCGCCGGCGCCGGCCGCCGCTGCCGACAAGTCTCGGCATCCGATCTCCTCCATCCTCC
GDS0252	GTCTCATCGCTGTTCACTGTT	CGAGCAAACCAACCATCCAA	178	GTCTCATCGCTGTTCACTGTTTTTTGTAGCTTATTCGACTTAGCGTCAATATAAGTATTCCTTATTATGTTACTCTCTCTCTCTCTCTCTCTCTTGTGTTTGAGAAATGGATTGTTGAATGATTGACTGTATGATTGCAGCTTCGTCTGTGATGCTTTGGATGGTTGGTTTGCTCG
GDS0134	CCTCCGATTTCTCAGACC	CATGGACCACATCATTGC	273	CCTCCGATTTCTCAGACCGATCCGGTCAGTGATCTACAGTCCCGGTTCGGGACGATTCATGACCGGTTGTTTTATTTTTATATATTTTACTGCTCAATAAACTTGAAATTTTAGTTGGTGATAACGTGGACCATGCAATGATCTAATTGTACGACTGCACAATATATATATATATATATATATATATAGATCTTCTATTCTATGCAGATAAGCTTGAAGTTCAATAATGCTTGCTTGTGCAATCGGTAAATATACGCAATGATGTGGTCCATG

Markers	Forward primer(5'-3')	Reverse primer(3'-5')	Product size(bp)	SSR containing sequence
GDS0615	CCACATACA GACTACAAC TTAC	GGAGGAGC GTATTATGA ACA	230	CCACATACAGACTACAACTTACAAGAACTTCCCACAAATACCCTTTTCTTC CGTAGATGGAATTCCTTTCATTATTTTCCGTTTTCTCCCATTCAAAGCTTCA ACAAAGAGAGAGATAGAGATAGAGAGAGGGAGAGAGAGAGAGAGAGAG AAAATCTCCATACCCCAGTCCTTACCCTTTTCTTCTTTCTCTCTCTGCATAA GAGTGAGTGTTCATAATACGCTCCTCC
GDS1059	TATAACCCG TATATCCCT ACCC	CTGGCACA CATATCATA TTTGG	185	TATAACCCGTATATCCCTACCCCTCTACTCTTCTAGATTGTTCTGCTCATAT ATATATATATATATAATATATGCTTGTACGTATTAGTGTGAGCATGTGGGG TATTGAAGATATTTTGTGGGAGGGAGGTATACCCCCATACCCATACGTAT CCACCGCATCCCAAATATGATATGTGTGCCAG
GDS0997	GGTGAGAT GCCATTATC TGACT	GAGAGTTA CAGTTCCA GCACTT	278	GGTGAGATGCCATTATCTGACTTCGAGAACCCTCTTTCATCGAGACCCACC TCTTTATTCAACCTTACTTCCAATCCCAATCCCGTTTCATTCCCGCGCACAA ATTACAATCCCCTTTTCTTCTTCTTCTTCTTCTTTCCATATACAAGAAG CTTCTCCTTCTAGTGTGTTCCAATCCCAAATTCATTTGATCGGAGGAAACA TATGTTTCCAAACTGGGAAGTTGTTCTGCTCAGTAGCTGCGGAGAAATTTA AGTGCTGGAACTGTAACTCTC

Markers	Forward primer(5'-3')	Reverse primer(3'-5')	Product size (bp)	SSR containing sequence
IB 1809	CTTCTCTTG CTCGCCTGT TC	GATAGTCG GAGGCATC TCCA	144-155	GCTTCTCTTGCTCGCCTGTTCCTACCGGAAACGCCCCTCCTCCTCCTCCTCC TCCTCCTCCTCCTCCGCCGCCGCCGCCGCCGCCGATCCCGATGAGAAATCC AGCCATTCCGCCGCCAACAGA
IB-242	GCGGAACG GACGAGAA AA	ATGGCAGA GTGAAAAT GGAACA	105-142	CATGGCAGAGTGAAAATGGAACAAATATCAGATTTCACATACAAACTGTG ATCTGCGCAGGCGTCCTTGCTCGCTCTCAATCTCTCAATCTCTCTCTCTCTC TCTCTCACTCTCTCT

Table 6. Summary of available SSR markers linked to β-carotene

3.2.3. Sequence similarity searching and analysis

The available EST-SSR marker sequences that flanks the QTL region are taken to search the similarity with the genome assembly of sweet potato provided by sweet potato genome browser in ipomoea genome hub database and sweet potato genomics resource database. The genome browser consists of the BLAT tool which will rapidly locate the position by homology alignment, provided that the region has been sequenced. BLAT (BLAST-Like Alignment Tool) is a very fast sequence alignment tool similar to BLAST. On DNA queries, BLAT is designed to quickly find sequences with 95% or greater similarity of length 25 bases or more. It may miss genomic alignments that are more divergent or shorter than these minimums, although it will find perfect sequence matches of 33 bases and sometimes as few as 22 bases. The tool is capable of aligning sequences that contain large introns. On protein queries, BLAT rapidly locates genomic sequences with 80% or greater similarity of length 20 amino acids or more.

3.2.3.1. Making a BLAT query

To locate a nucleotide or protein within a genome using BLAT:

1. Open the BLAT Search Genome page by clicking the BLAT link on the top menu bar of any of the Genome Browser pages.
2. Select the genome, assembly, query type, output sort order, and output type. To order the search results based on the closeness of the sequence match, choose one of the score options in the *Sort output* menu. The score is determined by the number of matches vs. mismatches in the final alignment of the query to the genome
3. If the sequence to be uploaded is in an unformatted plain text file, enter the file name in the *Upload sequence* text box, then click the *submit file* button. Otherwise, paste the sequence or fasta-formatted list into the large edit box, and then click the *submit* button. Input sequence can be obtained from the Genome Browser as well as from a custom annotation track.

The six EST-SSR markers, GDS0215, GDS1059, GDS0134, GDS0997, GDS0252 and GDS0615 sequences were aligned with the sweet potato genome assembly to locate the position of QTL on the genome. The sequence of SSR markers (Buteler *et al.*, 1999) that retrieved after the ePCR were searched for similarity against the sweet potato genome assembly in *Ipomoea batatas* genome browser. This was used to identify the chromosomal locations of the QTL by the alignment of these flanking markers to the genome. The chromosomal location of the identified QTLs were downloaded from the *Ipomoea batatas* genome browser. Chromosomal regions spanning the positions of flanking markers were investigated for candidate genes based on annotated biochemical functions. The candidate gene was identified from the thus determined chromosomal region by using the OmicsBox software. The main steps involved in OmicsBox to identify the candidate gene and its functional annotation are:

1. **Loading sequences**: The Chromosomal location sequences of QTL downloaded from *Ipomoea batatas* genome browser in FASTA format were loaded.
2. **Blast**: OmicsBox uses the Basic Local Alignment Search Tool (BLAST) to find sequences similar to the query set. BLAST can be performed in different actions:

- CloudBlast: This is a feature for massive sequence alignment tasks. CloudBlast is a high-performance, secure and cost-optimized solution for the analysis. This service is totally independent from the NCBI servers and provides fast and reliable sequence alignments.
- QBlast@NCBI: NCBI offers a public service that allows searching molecular sequence databases with the BLAST algorithm.
- Local BLAST against own database: It is used to BLAST to query a local/own database.

The parameters used in blast:

Blast Configuration Page

- e-mail address in case of using the NCBI BLAST web service
- BLAST program: blastx-fast
- BLAST DB (Database): nr
- Taxonomy Filter: No filter
- BLAST expect value: 1.0E-3
- Number of BLAST hits: 20

Advanced Configuration Page

Blast parameters

- Word size: 6
- Low complexity filter
- Filter Options:
- HSP length cutoff: 33
- HSP-Hit Coverage: 0

The results of the BLAST queries can also be directly saved to a file in different formats of xml, txt, html.

3. Gene Ontology Mapping

Mapping is the process of retrieving GO terms associated to the Hits obtained by the BLAST search. BLAST result accessions are used to retrieve gene names or symbols making use of two mapping files provided by the NCBI. Identified gene names are then searched in the species specific entries of the gene-product table of the GO database. Gene Bank identifiers (gi), the primary blast Hit ids, are used to retrieve UniProt IDs making use of a mapping file

from PIR (Non-redundant Reference Protein Database) including PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept and PDB.

4. Gene Ontology Annotation

This is the process of selecting GO terms from the GO pool obtained by the Mapping step and assigning them to the query sequences. The default parameters used were:

- Annotation Cut-Off (threshold): 55
- GO-Weight:5
- E-Value-Hit-Filter: 1.0E-6
- Hsp-HitCoverage CutOff: 0
- Hit Filter: 500

3.3. EXPERIMENTAL VALIDATION

The experimental validation of computationally predicted candidate genes were conducted by randomly choosing the predicted candidate genes geranyl geranyl diphosphate synthase and phytoene synthase genes involved in the carotenoid biosynthesis pathway. RT-qPCR was performed as described as below using the total RNA isolated from the tuber samples of two different varieties of sweet potato available at ICAR-CTCRI.

3.3.1. Selected sweet potato varieties

- Co-34: White fleshed tuber variety
- Bhu-sona: Orange fleshed tuber variety

3.3.2. Primer designing for the predicted genes

Primer3plus is a widely used program for designing PCR primers. PCR is an essential and ubiquitous tool in genetics and molecular biology. Primer3 can also design hybridization probes and sequencing primers. Primer3 picks primers for PCR reactions, considering certain criteria such as oligonucleotide melting temperature, size, GC content, primer dimer possibilities, PCR product size, positional constraints within the source sequence. The parameters considered in primer designing:

- Primer Length

It is generally accepted that the optimal length of PCR primers is 18-22 bp. This length is long enough for adequate specificity and short enough for primers to bind easily to the template at the annealing temperature.

- Primer Melting Temperature

Primer melting temperature (T_m) by definition is the temperature at which one half of the DNA duplex will dissociate to become single stranded and indicates the duplex stability. Primers with melting temperatures in the range of 52-58°C generally produce the best results.

- GC content

The GC content (the number of G's and C's in the primer as a percentage of the total bases) of primer should be 40-60%.

- GC Clamp

The presence of G or C bases within the last five bases from the 3' end of the primers (GC Clamp) helps promote specific binding at the 3' end due to the stronger bonding of G and C bases. More than 3 G's or C's should be avoided in the last 5 bases at the 3'end of the primer.

3.3.3. RNA isolation

Total RNA isolation was performed from tubers of two different varieties of sweet potato such as Co-34 and Bhu-sona available at ICAR-CTCRI using RNeasy plant mini kit of Qiagen in accordance with manufacturer's protocol and lithium chloride method.

3.3.3.1. Procedure for Lithium chloride method

RNA was extracted from two tuber varieties Co-34 (White fleshed) and Bhu-sona (Orange fleshed) available at ICAR-CTCRI using Lithium chloride method. 100 mg tubers of both varieties were grinded with liquid nitrogen and 1mL of CTAB buffer (pre-warmed at 65°C for 10 minutes) was added to it. The extract is transferred to a fresh centrifuge tube and is centrifuged at 15000 rpm for 15 minutes. The supernatant was transferred to fresh tube and equal volume of chloroform: isoamyl (24:1) alcohol was added and centrifuged at 20,000g for 10 minutes at 4°C. The Chroroform: Isoamyl step was repeated. The upper layer was transferred to fresh tube and 0.25V of ice cold 10M lithium chloride was added, mixed well and incubated overnight at 20°C. On the next day it was centrifuged at 30,000g for 30 minutes at 4°C. The pellet was washed with 75% ethanol by centrifuging at 10,000 rpm at 4°C for 10 minutes. The washing step was repeated. RNA pellet was then dried at 37°C for 30 minutes and dissolved in 20µL RNase free water. The integrity of the RNA was verified by 1%

agarose gel electrophoresis and was viewed under gel documentation system. The isolated RNA was stored at -80°C.

3.3.4. RNA quantification

The concentration of RNA was determined by using a Nano-drop (using 1 OD260=40µg RNA). A260/A280 ratios were also calculated for each sample. The cDNA of the samples was prepared using Revert-Aid First strand cDNA synthesis kit of Thermo Fisher Scientific in accordance with manufacturer's protocol.

3.3.5. RT-qPCR

Real-time quantitative polymerase chain reaction (RT-qPCR) is a sensitive technique for gene expression studies. The qPCR reaction was performed with forward and reverse primers (specific to predicted candidate genes in carotenoid pathway GGPS and phytoene synthase) using cDNA samples from the tubers of two different sweet potato varieties Co-34 and Bhu-sona available at ICAR-CTCRI. 2X H-eff qPCR master mix, Roxwas used for qPCR amplification.

Table 7. RT-qPCR reaction profile

Components	Volume (µl)
Diluted cDNA	1.5 µl
Forward primer (F)	1 µl
Reverse primer (R)	1 µl
2X H-eff qPCR master mix, Rox (GCC Biotech)	5 µl
Double distilled water	1.5 µl

3.3.5.1 Thermal profile

- Initial denaturation - 95°C
- Denaturation - 95°C

- Annealing - 55°C
- Extension - 72°C

Number of cycles - 35-45 cycles, step 2-4

After the completion of the real time PCR reactions, the threshold cycle (Ct) was recorded and gene expression level was calculated using comparative Ct method or delta-delta Ct method

The relative gene expression level of white and orange fleshed sweet potato varieties is represented as $2^{-\Delta\Delta CT}$ method.

ΔCt= Ct (target gene) - Ct (reference gene)

$\Delta\Delta Ct$= ΔCt (sample) - Δ Ct (Control)

RESULTS

4. RESULTS

The main objective of the study was to identify the differentially expressed genes for various tuber colours in sweet potato using RNA sequenced data; to integrate QTL information on tuber colour with genomic information in sweet potato and to validate the identified candidate genes.

4.1. IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES FOR VARIOUS TUBER COLOURS IN SWEET POTATO

The preliminary transcriptome data of six tuber varieties and its replicates sequenced using Nextseq 500 platform was obtained from ICAR-CTCRI.

4.1.1. Raw sequence processing and quality control

The raw sequence data of the six tuber samples were analysed for their quality using FastQC software. The output from FastQC (v 0.11.7), after analyzing file of sequence reads, contains a flag of "PASS", "WARN" or "FAIL" assigned to each of the parameters including basic statistics, per base sequence quality, per sequence quality scores, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences and kmer content. The thresholds used to assign the flags are based on specific assumptions. Among all the tuber varieties and its replicons, six out of the eleven quality modules examined a warn or fail sign. These inadvertences appeared on per base sequence content, per base GC content, sequence length distribution, sequence duplication levels, overrepresented sequences and kmer content. The quality module per base sequence content in all tuber varieties showed a biased sequence composition, normally at the start of the read. This bias does not concern an absolute sequence but instead produce overrepresented sequences. This may be related with a problem in the library generation or a consequence of abnormal sequencing methods. The per base GC content module instead of focusing on all the four nucleotides, only focuses the G and C content. Sequence length distribution bias arises due to varying length of sequence fragments. It may be because of the

sequencing platforms used. The sequence duplication level in all tubers showed a high level of duplication which is more likely to indicate some kind of enrichment biases. The overrepresentation module showed sequences are very much overrepresented indicating that the library is contaminated. This module will often be triggered when used to analyse RNA libraries where sequences are not subjected to random fragmentation, and the same sequence may naturally present in a significant proportion of the library. All the tuber libraries showed Kmer bias at the start of the read, may be due to an incomplete sampling of the possible random primers. Based on FastQC analysis, the bases with low sequencing quality should be trimmed ensuring the quality of the high throughput data.

4.1.2. Quality trimming and adapter removal

The sequenced raw data after quality checking was processed to obtain high quality clean reads using TRIMMOMATIC. The raw data are analysed to remove the adapter contamination, ambiguous reads (reads with unknown nucleotides "N") and low quality sequences *i.e.,* reads with more than 10% quality threshold. After the trimming process of raw data, high quality reads were retained for all the six tuber varieties and its replicates (Table 8). This obtained high quality reads were used for *de novo* transcriptome assembly.

Table 8. Total number of sequences obtained after trimming

Library	Total sequences before trimming	Total sequences after trimming
Tuber1_R1	26100791	26081266
Tuber1_R2	26100791	26060792
Tuber2_R1	23454234	23436464
Tuber2_R1	23454234	23416026
Tuber3_R1	24525000	24506748
Tuber3_R2	24525000	24483941
Tuber4_R1	23138799	23121241
Tuber4_R2	23138799	23104005
Tuber5_R1	23853281	23835350
Tuber5_R2	23853281	23816340
Tuber6_R1	25860443	25841424
Tuber6_R2	25860443	25841424

4.1.3. *De novo* transcriptome assembly

The filtered high quality reads of the six tuber samples and its replicates were assembled into transcripts using Trinity RNA-Seq assembler (v2.8.5) with default parameters of kmer. The assembly finally produced 111231 transcripts with the mean size of 532 bp and an N50 of 1324 for Tuber 1_R1, 110299 transcripts with the mean size of 531 bp and an N50 of 1329 for Tuber1_R2, 105694 transcripts with the mean size of 543 bp and an N50 of 1362 for Tuber2R1, 103338 transcripts with the mean size of 536 bp and an N50 of 1356 for Tuber2R2, 132409 transcripts with the mean size of 566 bp and an N50 of 1456 for Tuber3_R1, 128902 transcripts with the mean size of 567 bp and an N50 of 1474 for Tuber3_R2, 121941 transcripts with the mean size of 527 bp and an N50 of 1327 for Tuber4_R1, 121263 transcripts with the mean size of 523bp and an N50 of 1345 for Tuber 4_R2, 106347 transcripts with the mean size of 537 bp and

anN50 of 1350 for Tuber5_R1, 103599 transcripts with the mean size of 528 bp and N50 of 1343 for Tuber5_R2, 108964 transcripts with the mean size of 545 bp and an N50 of 1358 for Tuber6_R1 and 106177 transcripts with the mean size of 541 bp and an N50 of 1374 for Tuber6_R2 (Table 9).

4.1.4. Transcript level quantification

Gene and isoform expression levels of each of the tubers transcripts sequences were obtained from the transcript level quantification in OmicsBox. The fastQ sequencing reads and FASTA file of transcript sequences generated from *de novo* assembly of each sample were used to obtain a count table representing the expression levels. The program mainly aligns the reads against the transcript reference sequences and calculate the relative abundances. The analysis returned two outputs of count tables showing both expression level of each transcript or isoform and genes. Library size per sample showed the number of read counts aligned to the transcriptome of each of the sample (Figure 2). Counts per category bar chart marked the number of reads of each input file sorted by different categories of aligned only once to a reference transcript, reads aligned more than one to the reference transcript and reads that have not been assigned to any of the transcript (Figure 3).

Table 9. Assembly results of the tuber transcriptome using Trinity

Sample	No. of transcripts	No. of genes	Total transcript length (bases)	N50	Mean transcript length
Tuber1R1	111231	61890	94063221	1324	532
Tuber1R2	110299	62357	93451604	1329	531
Tuber2R1	105694	59768	91969634	1362	543
Tuber2R2	103338	59857	89051980	1356	536
Tuber3R1	132409	67189	121255186	1456	566
Tuber3R2	128902	66251	118262930	1474	567
Tuber4R1	121941	66330	103304056	1327	527
Tuber4R2	121263	67406	103014201	1345	523
Tuber5R1	106347	59941	91341935	1350	537
Tuber5R2	103599	59768	88608751	1343	528
Tuber6R1	108964	59695	94573532	1358	545
Tuber6R2	106177	59440	92713212	1374	541

Table 10. Summary of transcript level quantification at isoform level

Input Reads		Aligned Reads		Total	
Name	Total Records	Aligned 1 Time	Aligned Multiple Times	Overall Alignment Rate	Not Aligned
Tuber4_R2	23,138,799	8,027,901 / 34.69%	11,338,667 / 49.00%	19,366,568 / 83.70%	3,772,231 / 16.30%
Tuber3_R1	24,525,000	9,457,816 / 38.56%	10,466,931 / 42.68%	19,924,747 / 81.24%	4,600,253 / 18.76%
Tuber4_R1	23,138,799	8,016,450 / 34.65%	11,393,410 / 49.24%	19,409,860 / 83.88%	3,728,939 / 16.12%
Tuber5_R2	23,853,281	8,035,229 / 33.69%	11,514,425 / 48.27%	19,549,654 / 81.96%	4,303,627 / 18.04%
Tuber2_R1	23,454,234	8,529,965 / 36.37%	11,797,667 / 50.30%	20,327,632 / 86.67%	3,126,602 / 13.33%
Tuber3_R2	24,525,000	9,381,170 / 38.25%	10,037,514 / 40.93%	19,418,684 / 79.18%	5,106,316 / 20.82%
Tuber2_R2	23,454,234	8,685,316 / 37.03%	11,214,618 / 47.81%	19,899,934 / 84.85%	3,554,300 / 15.15%
Tuber1_R1	26,100,791	9,642,562 / 36.94%	13,053,708 / 50.01%	22,696,270 / 86.96%	3,404,521 / 13.04%
Tuber1_R2	26,100,791	9,682,970 / 37.10%	12,678,328 / 48.57%	22,361,298 / 85.67%	3,739,493 / 14.33%
Tuber6_R1	25,860,443	8,924,819 / 34.51%	12,856,881 / 49.72%	21,781,700 / 84.23%	4,078,743 / 15.77%
Tuber6_R2	25,860,443	8,928,470 / 34.53%	12,349,968 / 47.76%	21,278,438 / 82.28%	4,582,005 / 17.72%
Tuber5_R1	23,853,281	8,059,032 / 33.79%	11,912,929 / 49.94%	19,971,961 / 83.73%	3,881,320 / 16.27%

Figure 2. Bar chart showing the number of read counts aligned to transcriptome features contained in each sample.

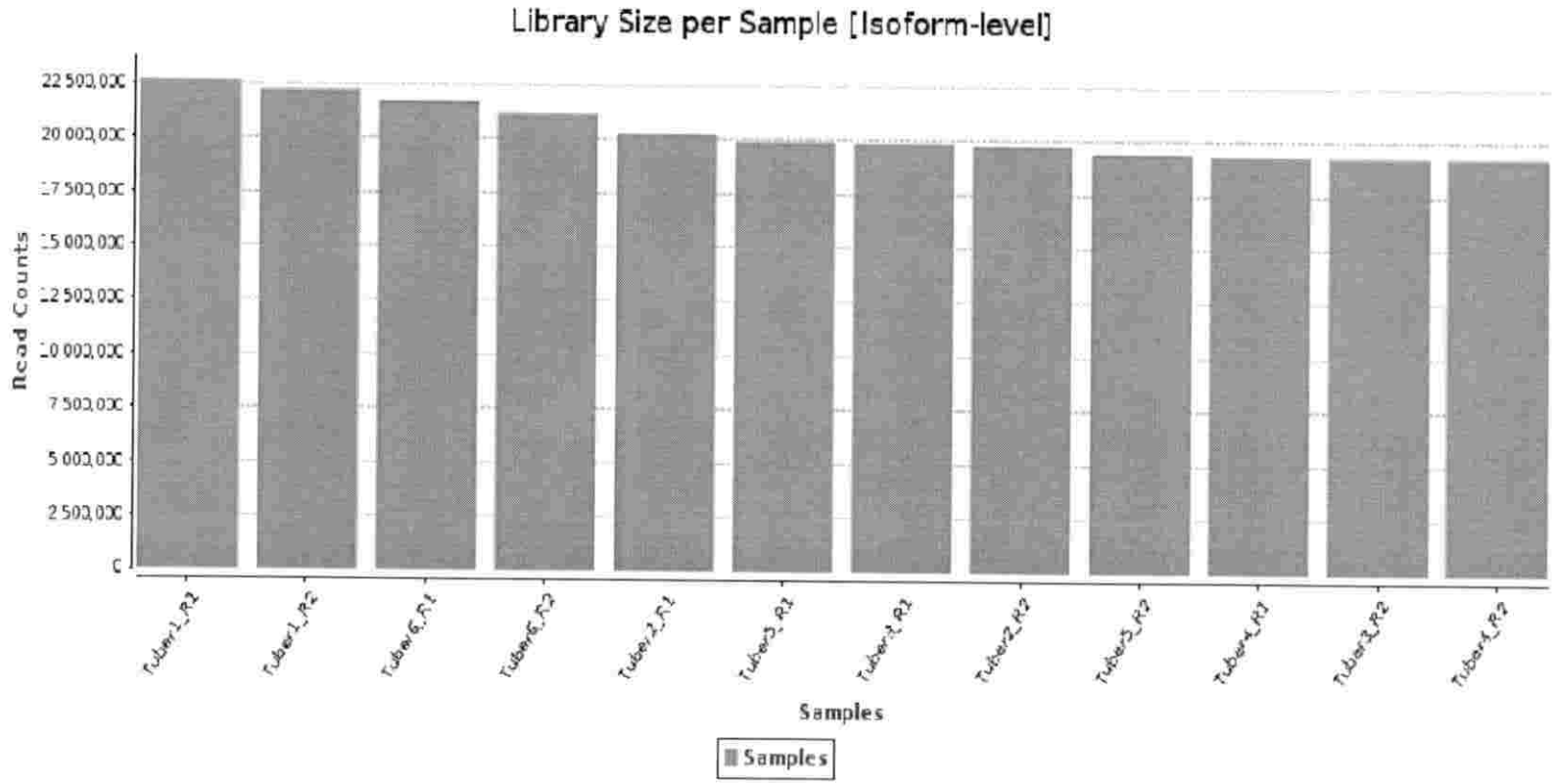
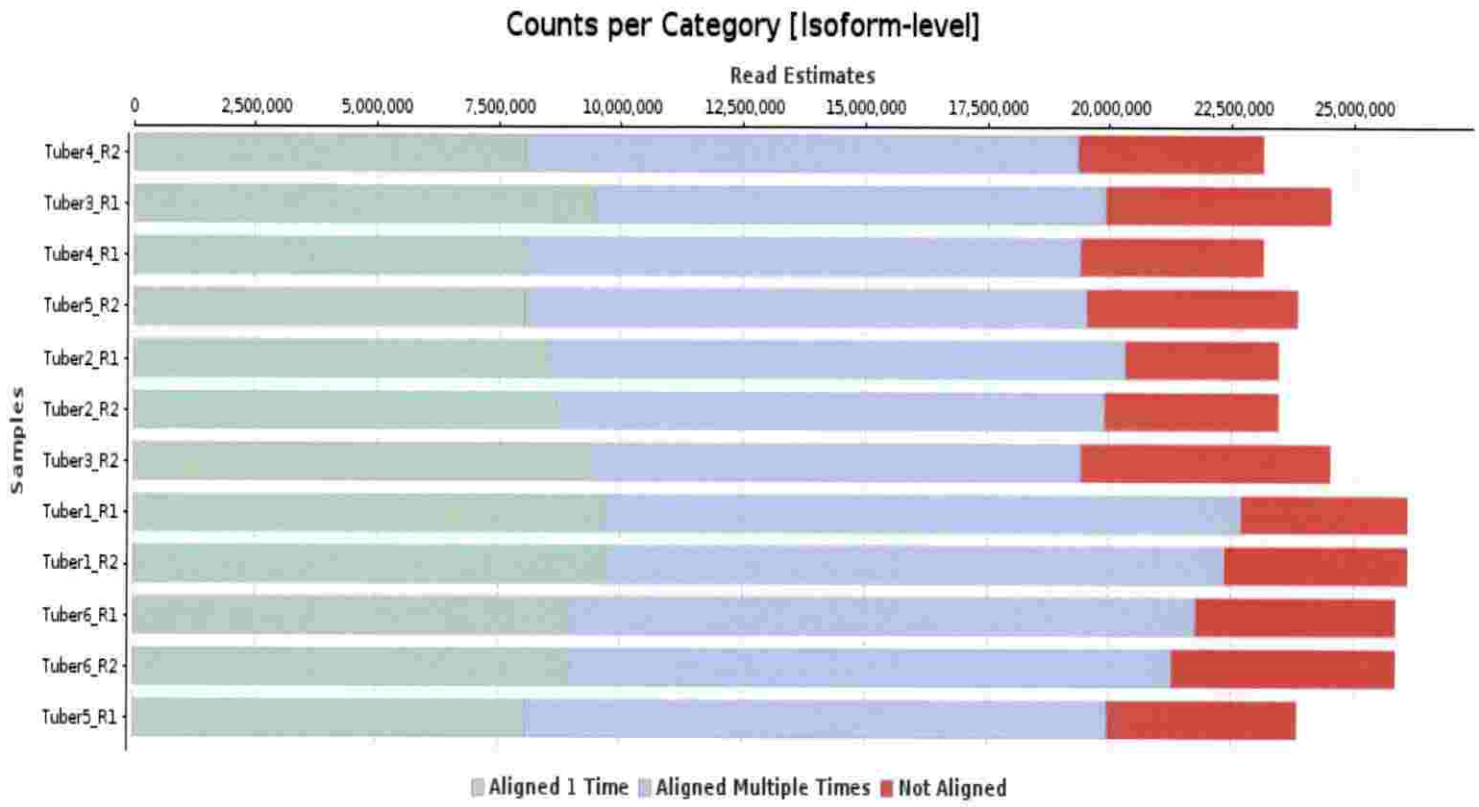


Figure 3. Bar chart showing the number of reads of each input file sorted by different categories



4.1.5. Pairwise Differential expression analysis

Pairwise differential expression analysis for the three different tubers colours were performed. OmicsBox allowed the identification of differentially expressed genes considering two different conditions from the analysis of count data. The analysis generated top Differentially Expressed Genes tags in a data for a pair of groups ranked by adjusted p-value. For the differential expression analysis pair of groups taken were orange and white, orange and purple, purple and white in which all the former ones as the contrasting conditions and latter one as the reference. The genes that are upregulated (+1) or downregulated (-1) were counted with a parameter of FDR<0.05 and logFC $\geq$ 1Based on the RNA seq data of white (Tuber1 and Tuber2) and orange (Tuber5 and Tuber6) 22,534 out of the 111,231 transcripts were variably expressed between white and orange varieties. Among them, 5472 were upregulated and 17,062 were downregulated in orange compared to white (Table 11). From orange and purple (Tuber3 and Tuber4) RNA seq data, 27,431 out of the 111,231 transcripts were found to be differentially expressed between orange and purple varieties with 11,670 upregulated genes and 15,761 downregulated genes in orange compared to purple (Table 12). RNA seq data of purple and white produced 22,590 variably expressed genes out of 111,231 transcripts between purple and white varieties. Among them, 7,622 were upregulated and 14,968 were downregulated in purple compared to white. The analysis generated a summary of the differentially expressed genes (Table 13). The leading log fold change means the average of the largest absolute log fold changes between each pair of samples. MA plot representing relationship between logFC and average expression level for differentially expressed genes, highlightened in red, blue line indicating genes up or down regulated two fold (Figures 5, 9&13). This plot explains the quality of the reads. Volcano plot illustrates the native log of adjusted p-value (FDR) to the log of fold changes where green indicating the upregulated genes and red indicating down regulated genes(Figures6,10&14), heatmap showing visual representation of data in which numerical values of points are represented by a range of colours in which red indicating upregulated genes and green indicating downregulated ones (Figures

7,11&15). The dendrograms added to the left and top side are produced by a hierarchical clustering method that takes as input the distance computed between genes (left) and samples (top). Venn diagram represented 2656 upregulated genes in common between orange-purple and orange-white libraries, 1459 upregulated genes in common between orange-white and purple-white libraries, 58 upregulated genes in common between all the three libraries (Figure 16). In case of downregulated genes, 6325 were in common between orange-white and purple-white libraries, 7932 in common with orange-purple and orange-white libraries and 512 genes in common between all the three libraries (Figure 17).

Table 11. Summary of pairwise differential expression analysis between orange and white libraries

Dataset Overview

- Number of total features: **111,231**
- Number of filtered features: **742**
- Number of features after filtering: **110,489**
- Number of analyzed samples: **12/12**

Results

Number of differentially expressed (DE) features (FDR <= 0.05): **22,534**

- Up-regulated (Log FC >= 1): **5,472**
- Down-regulated (Log FC <= -1) : **17,062**

Experimental Design

Sample	Lib. size (pre-filter)	Lib. size (post-filter)	Norm. factor	Condition	Individual
Tuber1_R1	22,590,786	22,590,786	1.137	White	1
Tuber1_R2	22,153,179	22,153,179	1.020	White	1
Tuber2_R1	20,284,719	20,284,719	1.012	White	2
Tuber2_R2	18,755,141	18,755,141	0.871	White	2
Tuber3_R1	19,844,445	19,844,445	0.076	Purple	3
Tuber3_R2	19,274,848	19,274,848	0.920	Purple	3
Tuber4_R1	19,309,827	19,309,827	1.085	Purple	4
Tuber4_R2	19,263,017	19,263,017	1.065	Purple	4
Tuber5_R1	19,903,060	19,903,060	0.949	Orange	5
Tuber5_R2	19,471,455	19,471,455	0.905	Orange	5
Tuber6_R1	21,707,804	21,707,804	0.969	Orange	6
Tuber6_R2	21,120,308	21,120,308	0.967	Orange	6

Analysis Parameters

Parameter	Value
CPM Filter	0.0
Samples reaching CPM Filter	1
Normalization Method	TMM (Trimmed mean of M values)
Experimental Design File	/Users/bdelon/Desktop/Project -Sweet potato/pairwise RESULTS/Count table/experimental_design.txt
Design Type	Simple Design
Primary Experimental Factor	Condition
Primary Contrast Condition	Orange
Primary Reference Condition	White

Figure 4. Bar chart which showing the overall result of differential expression analysis between orange and white libraries

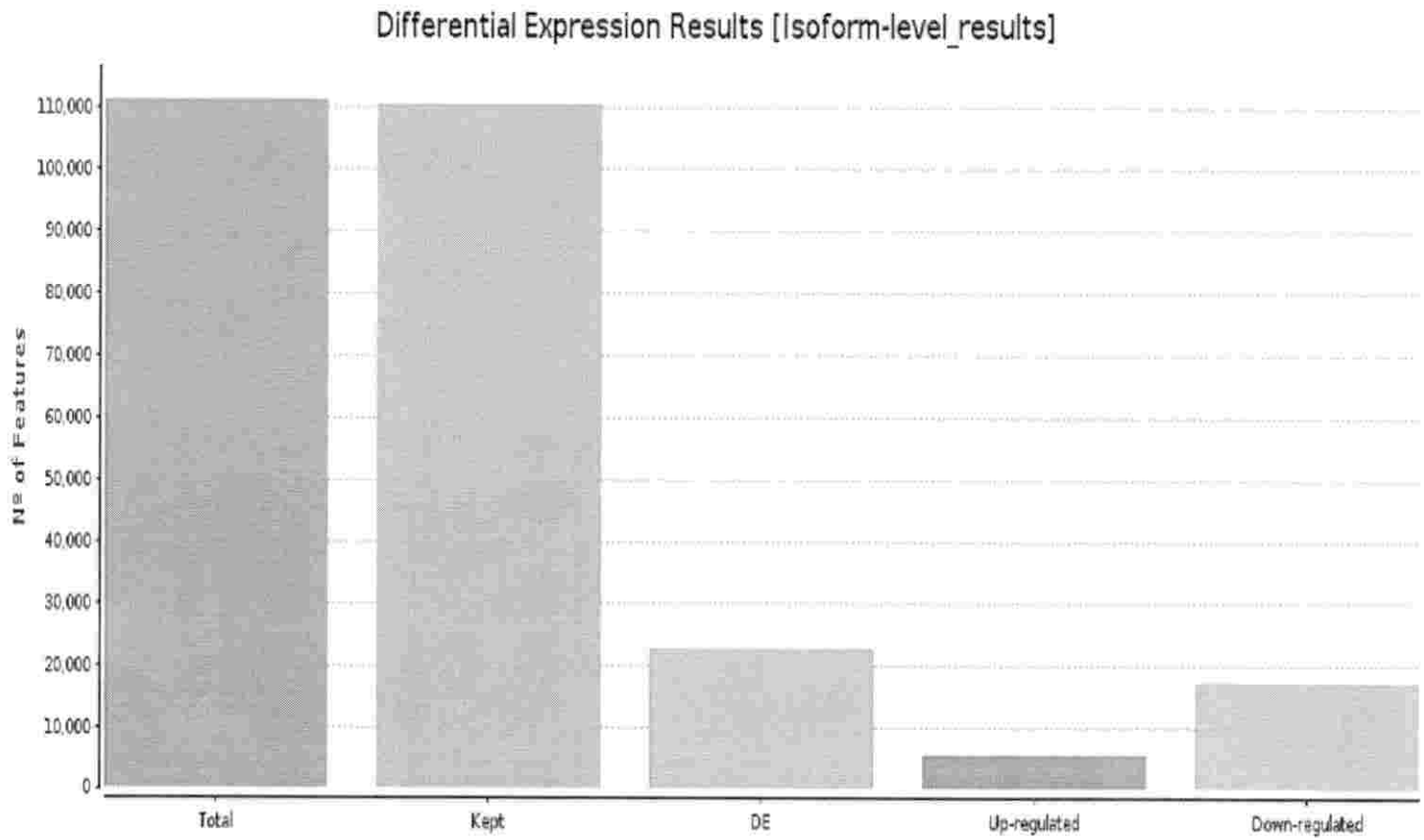


Figure 5. Scatter plot showing the log of the fold changes versus the average of the log of the CPM for orange and white pair analysis

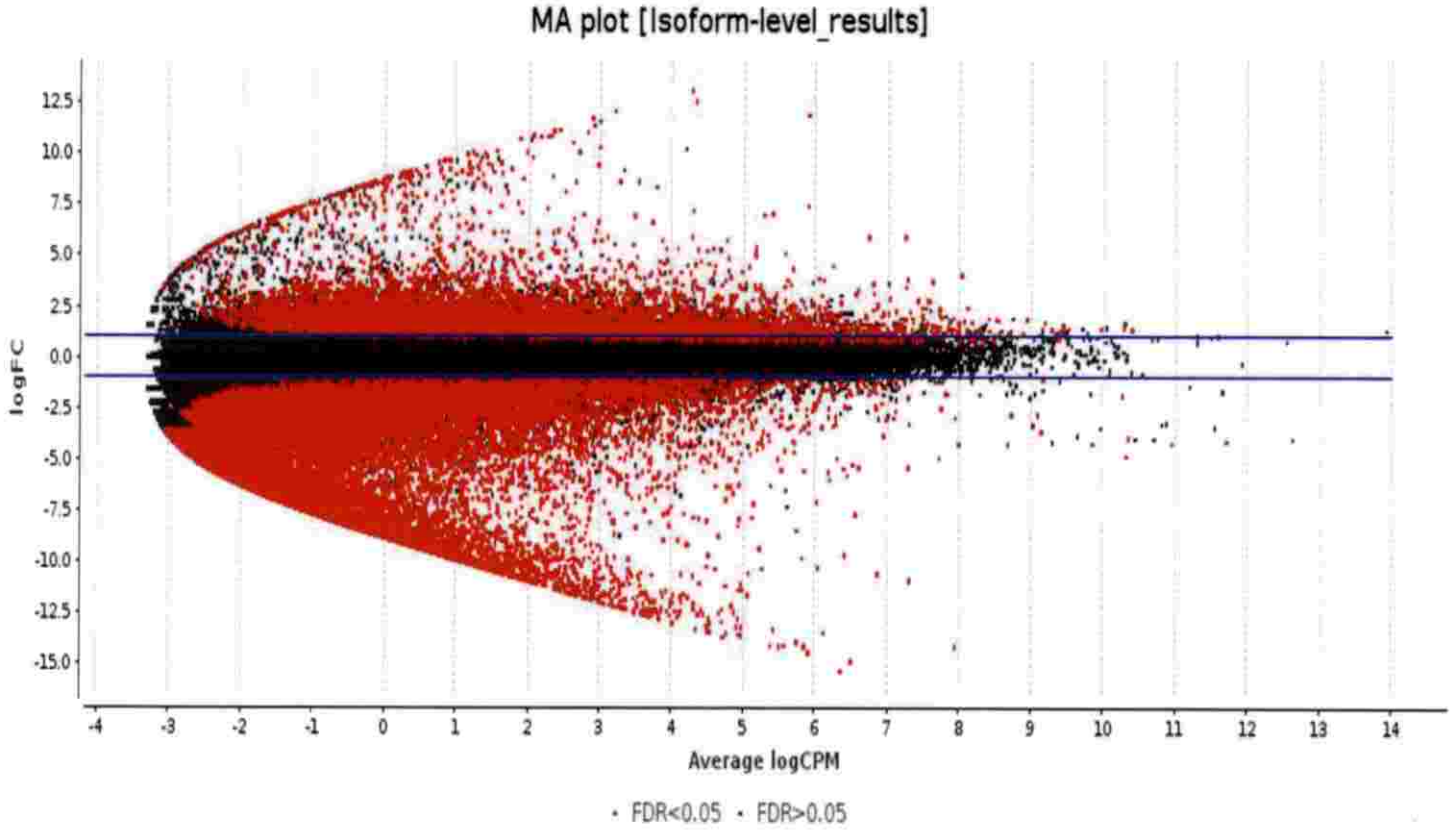


Figure 6. Scatter plot representing negative log of the FDR versus the log of the fold changes for orange and white pair analysis

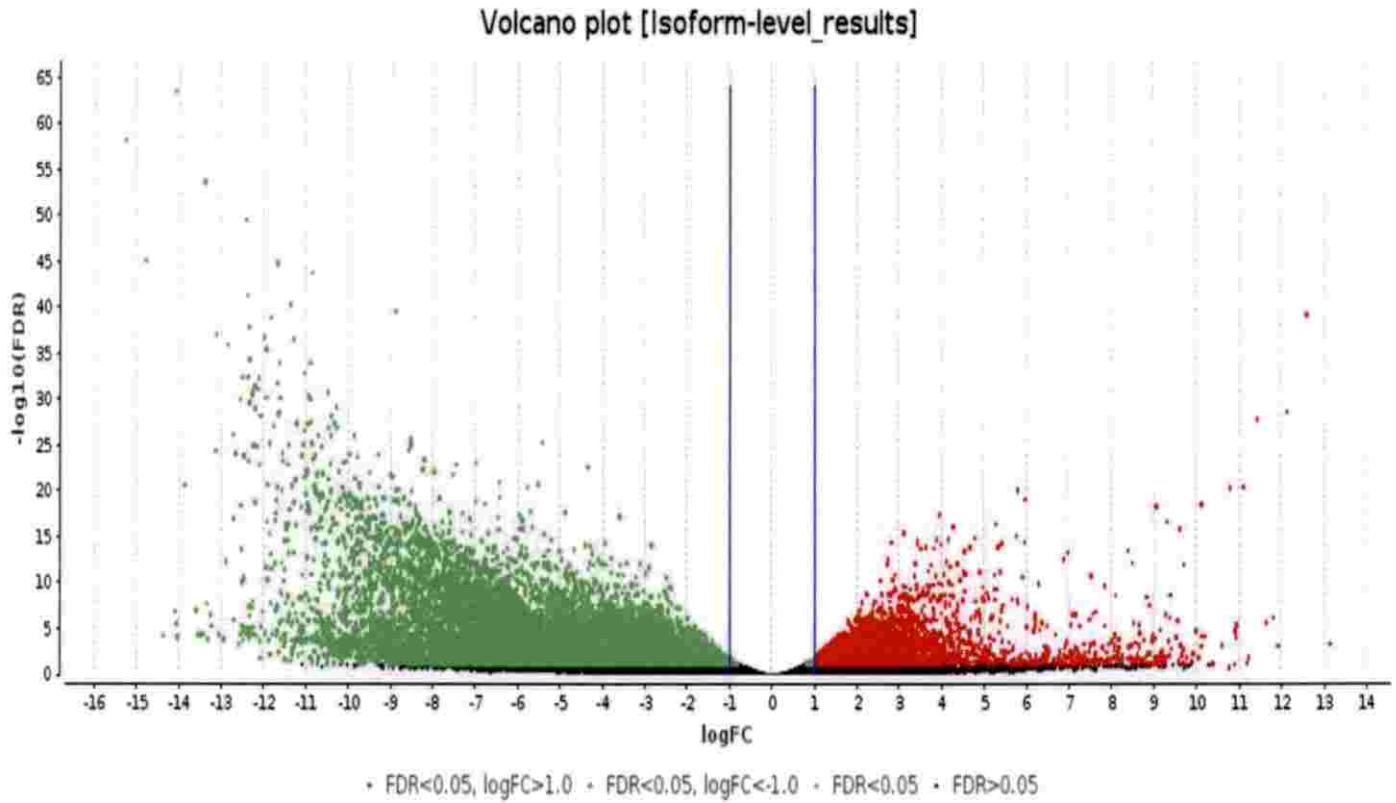


Figure 7. Heatmap showing the top 50 differentially expressed genes between orange and white pair analysis

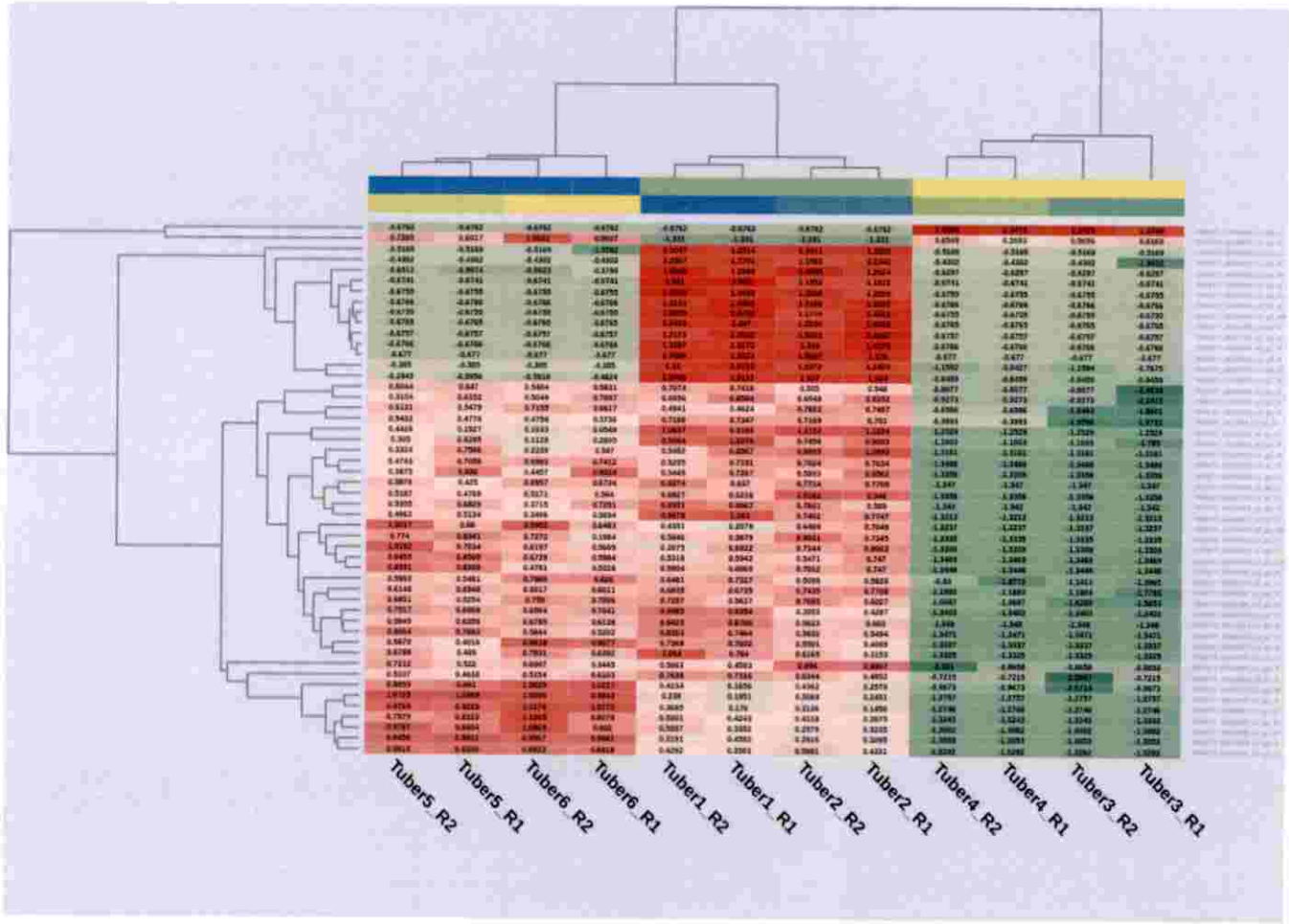


Table 12. Summary of pairwise differential expression analysis between orange and purple libraries

Dataset Overview

- Number of total features: **111,231**
- Number of filtered features: **742**
- Number of features after filtering: **110,489**
- Number of analyzed samples: **12/12**

Results

Number of differentially expressed (DE) features (FDR < 0.05): **27,431**

- Up-regulated (Log FC > 1): **11,670**
- Down-regulated (Log FC < -1) : **15,761**

Experimental Design

Sample	Lib. size (pre-filter)	Lib. size (post-filter)	Norm. factor	Condition	Individual
Tuber1_R1	22,590,786	22,590,786	1.137	White	1
Tuber1_R2	22,183,179	22,183,179	1.020	White	1
Tuber2_R1	20,234,712	20,234,712	1.012	White	2
Tuber2_R2	19,755,144	19,755,144	0.971	White	2
Tuber3_R1	19,844,445	19,844,445	0.976	Purple	3
Tuber3_R2	19,274,848	19,274,848	0.929	Purple	3
Tuber4_R1	19,309,827	19,309,827	1.095	Purple	4
Tuber4_R2	19,263,017	19,263,017	1.065	Purple	4
Tuber5_R1	19,903,060	19,903,060	0.949	Orange	5
Tuber5_R2	19,421,455	19,421,455	0.905	Orange	5
Tuber6_R1	21,707,804	21,707,804	0.998	Orange	6
Tuber6_R2	21,120,308	21,120,308	0.967	Orange	6

Analysis Parameters

Parameter	Value
CPM Filter	0.0
Samples reaching CPM Filter	1
Normalization Method	TMM (Trimmed mean of M values)
Experimental Design File	/home/bslctcri/Desktop/Project -Sweet potato/pairwise RESULTS/Count table /experimental_design.txt
Design Type	Simple Design
Primary Experimental Factor	Condition
Primary Contrast Condition	Orange
Primary Reference Condition	Purple

Figure 8. Bar chart which showing the overall result of differential expression analysis between orange and purple libraries

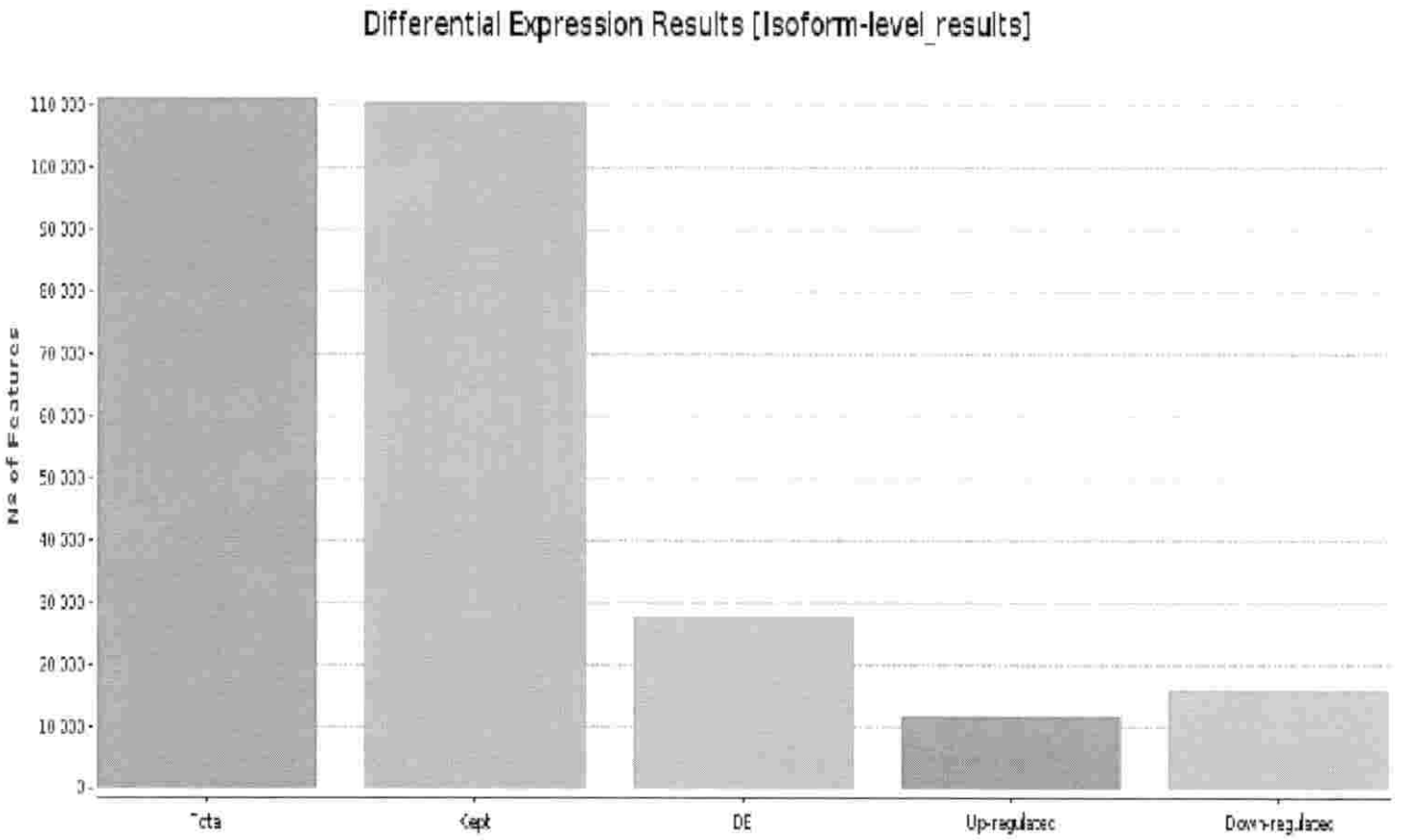


Figure 9. Scatter plot showing the log of the fold changes versus the average of the log of the CPM for orange and purple pair analysis

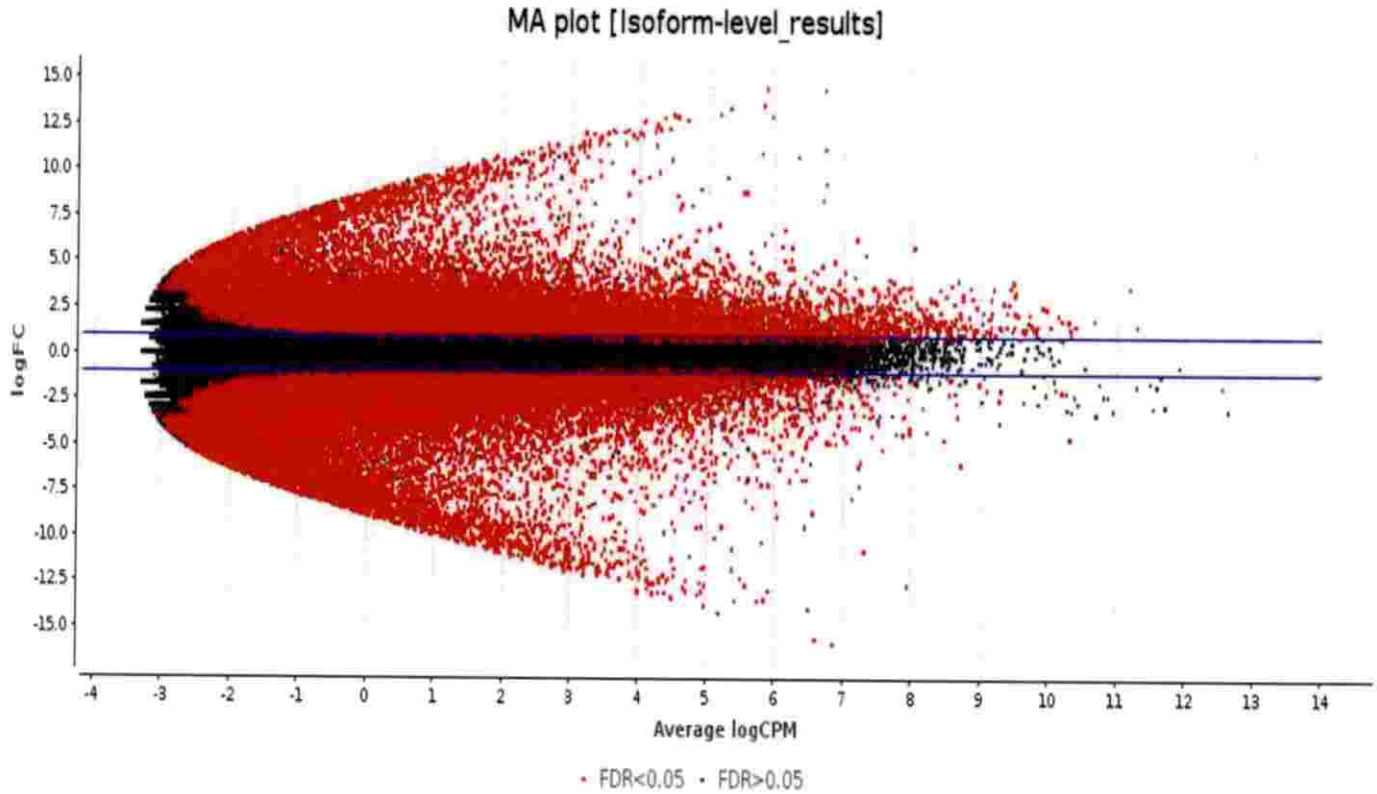


Figure 10. Scatter plot representing negative log of the FDR versus the log of the fold changes for orange and purple pair analysis

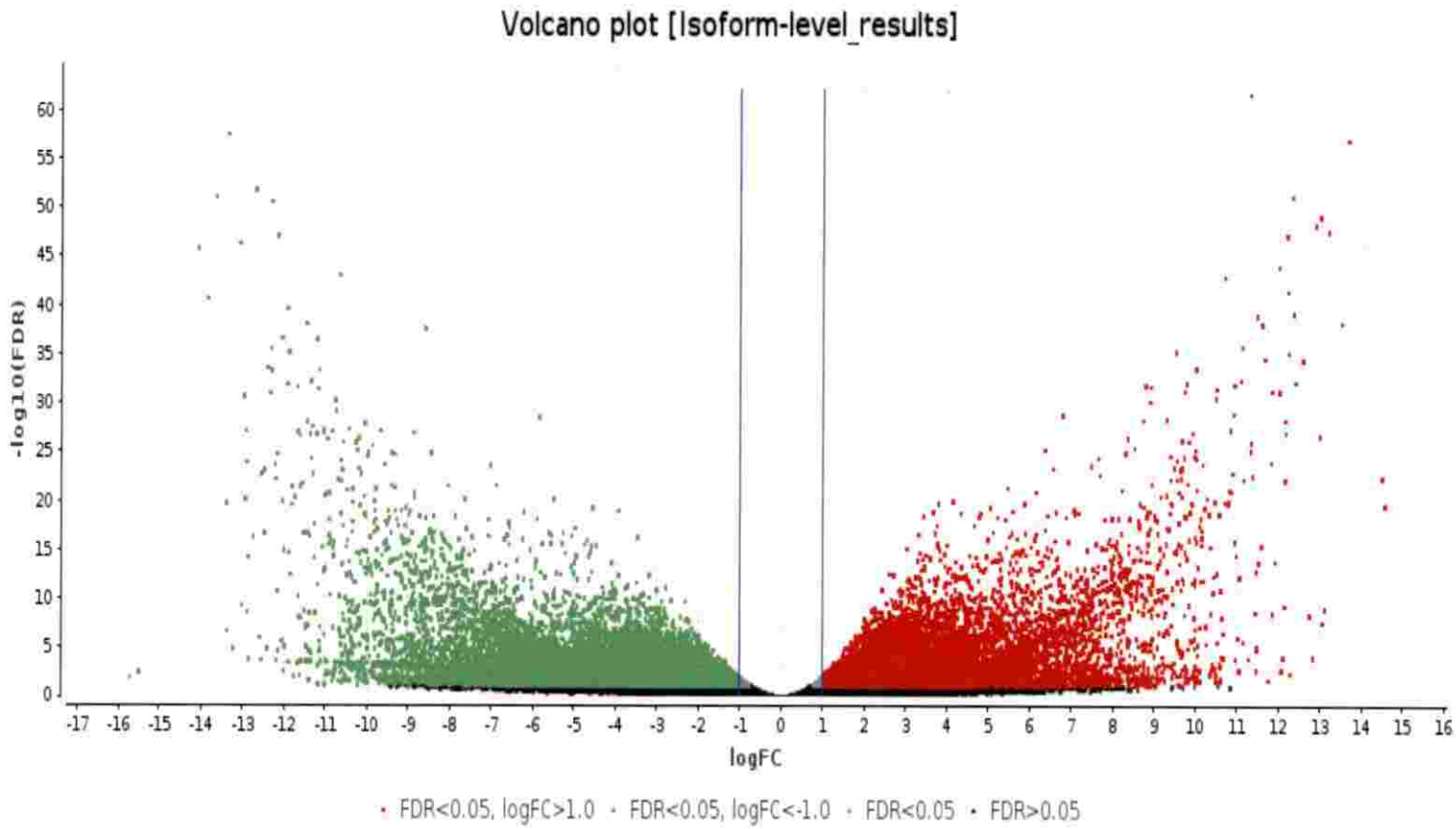


Figure 11. Heatmap showing the top 50 differentially expressed genes between orange and purple pair analysis

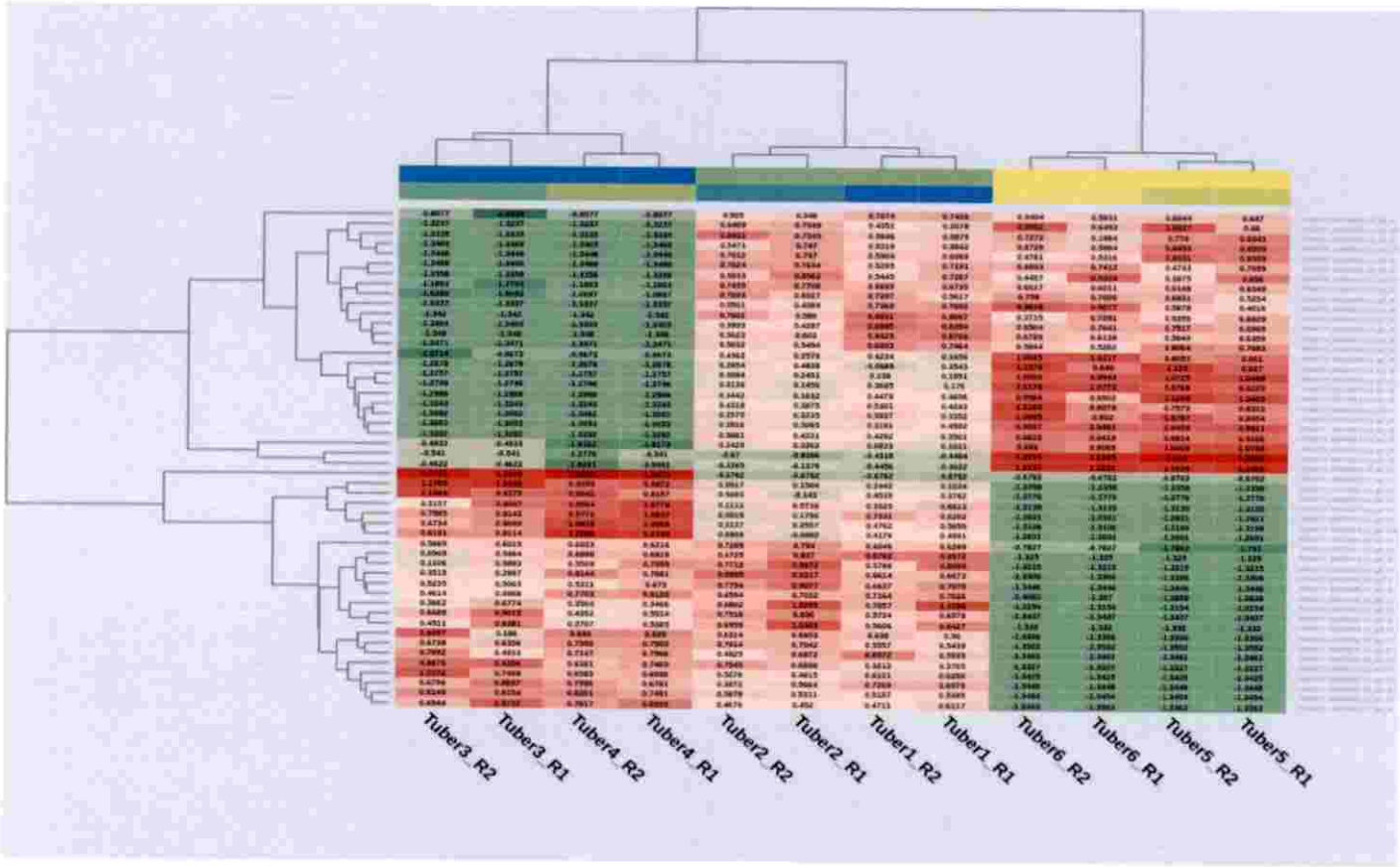


Table 13. Summary of pairwise differential expression analysis between purple and white libraries

Dataset Overview

- Number of total features: **111,231**
- Number of filtered features: **742**
- Number of features after filtering: **110,489**
- Number of analyzed samples: **12/12**

Results

Number of differentially expressed (DE) features (FDR < 0.05): **22,590**

- Up-regulated (Log FC > 1): **7,622**
- Down-regulated (Log FC < -1) : **14,968**

Experimental Design

Sample	Lib. size (pre-filter)	Lib. size (post-filter)	Norm. factor	Condition	Individual
Tuber1_R1	22,550,786	22,550,786	1.137	White	1
Tuber1_R2	22,183,179	22,183,179	1.020	White	1
Tuber2_R1	20,234,712	20,234,712	1.012	White	2
Tuber2_R2	19,755,144	19,755,144	0.971	White	2
Tuber3_R1	19,844,445	19,844,445	0.970	Purple	3
Tuber3_R2	19,274,848	19,274,848	0.929	Purple	3
Tuber4_R1	19,308,827	19,308,827	1.085	Purple	4
Tuber4_R2	19,263,017	19,263,017	1.065	Purple	4
Tuber5_R1	18,803,080	18,803,080	0.849	Orange	5
Tuber5_R2	19,421,455	19,421,455	0.905	Orange	5
Tuber6_R1	21,707,804	21,707,804	0.998	Orange	6
Tuber6_R2	21,120,308	21,120,308	0.987	Orange	6

Analysis Parameters

Parameter	Value
CPM Filter	0.0
Samples reaching CPM Filter	1
Normalization Method	TMM (Trimmed mean of M values)
Experimental Design File	/home/bskctcrl/Desktop/Project -Sweet potato/palrwise REGULTG/Count table /experimental_design.txt
Design Type	Simple Design
Primary Experimental Factor	Condition
Primary Contrast Condition	Purple
Primary Reference Condition	White

92

Figure 12. Bar chart which showing the overall result of differential expression analysis between purple and white libraries

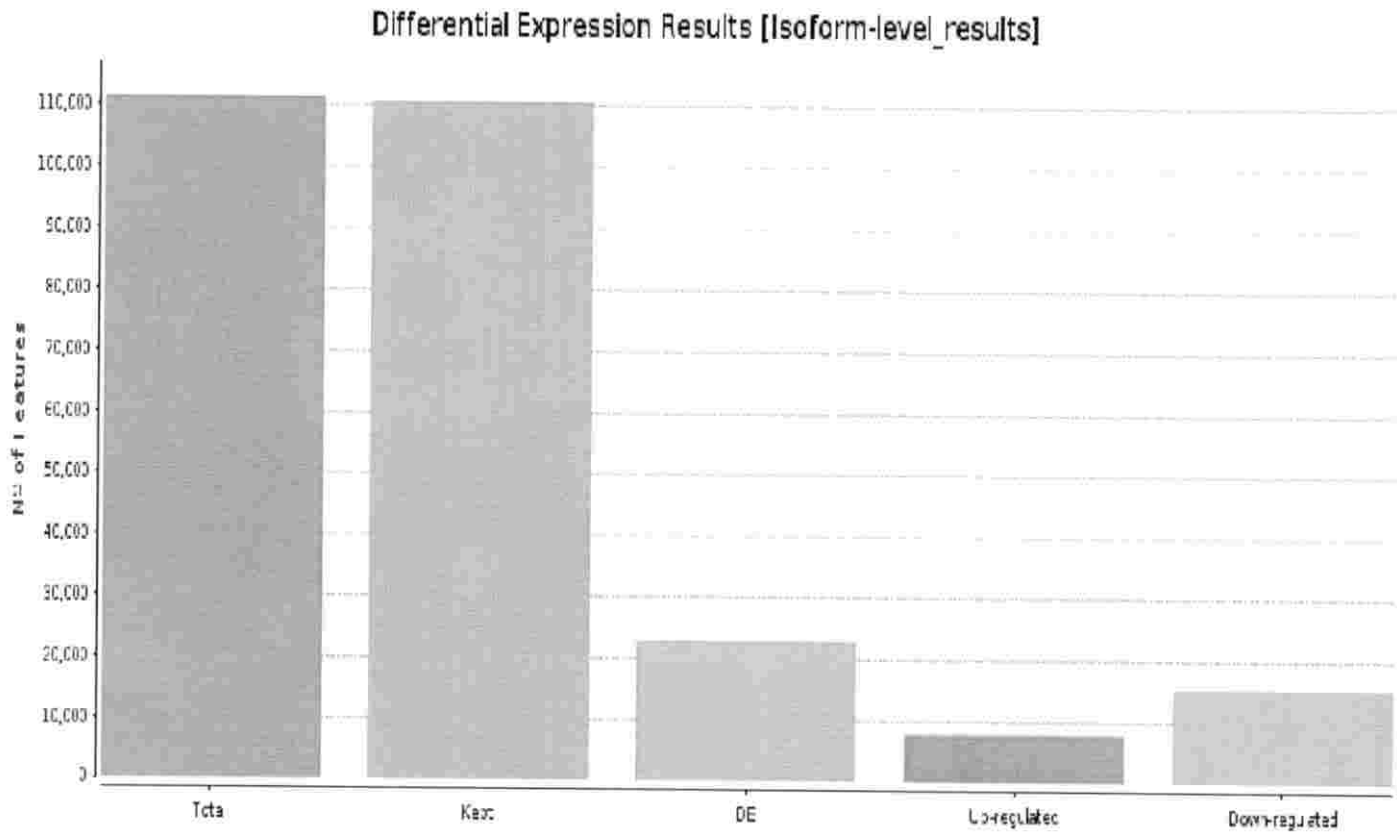


Figure 13. Scatter plot showing the log of the fold changes versus the average of the log of the CPM for purple and white pair analysis

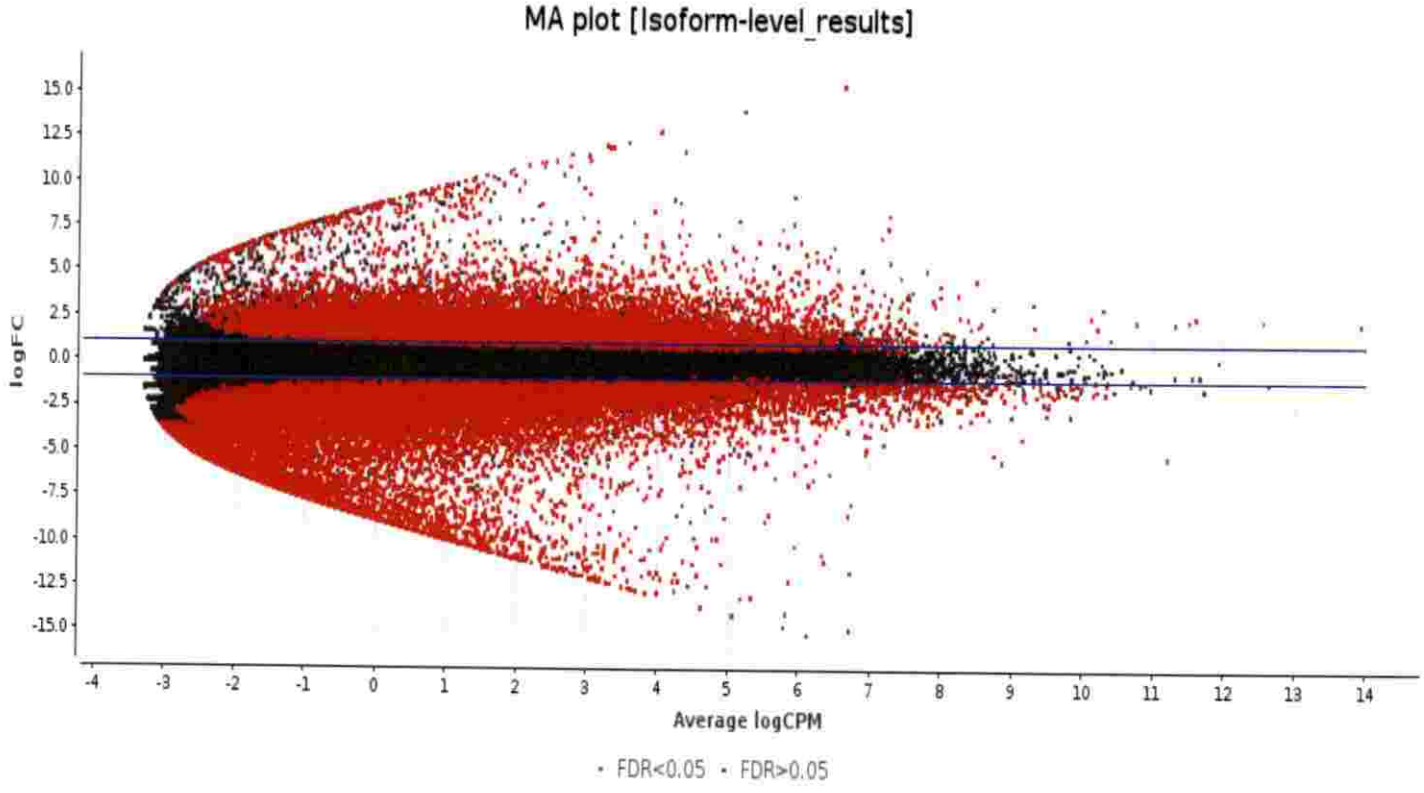


Figure 14. Scatter plot representing negative log of the FDR versus the log of the fold changes for orange and purple pair analysis

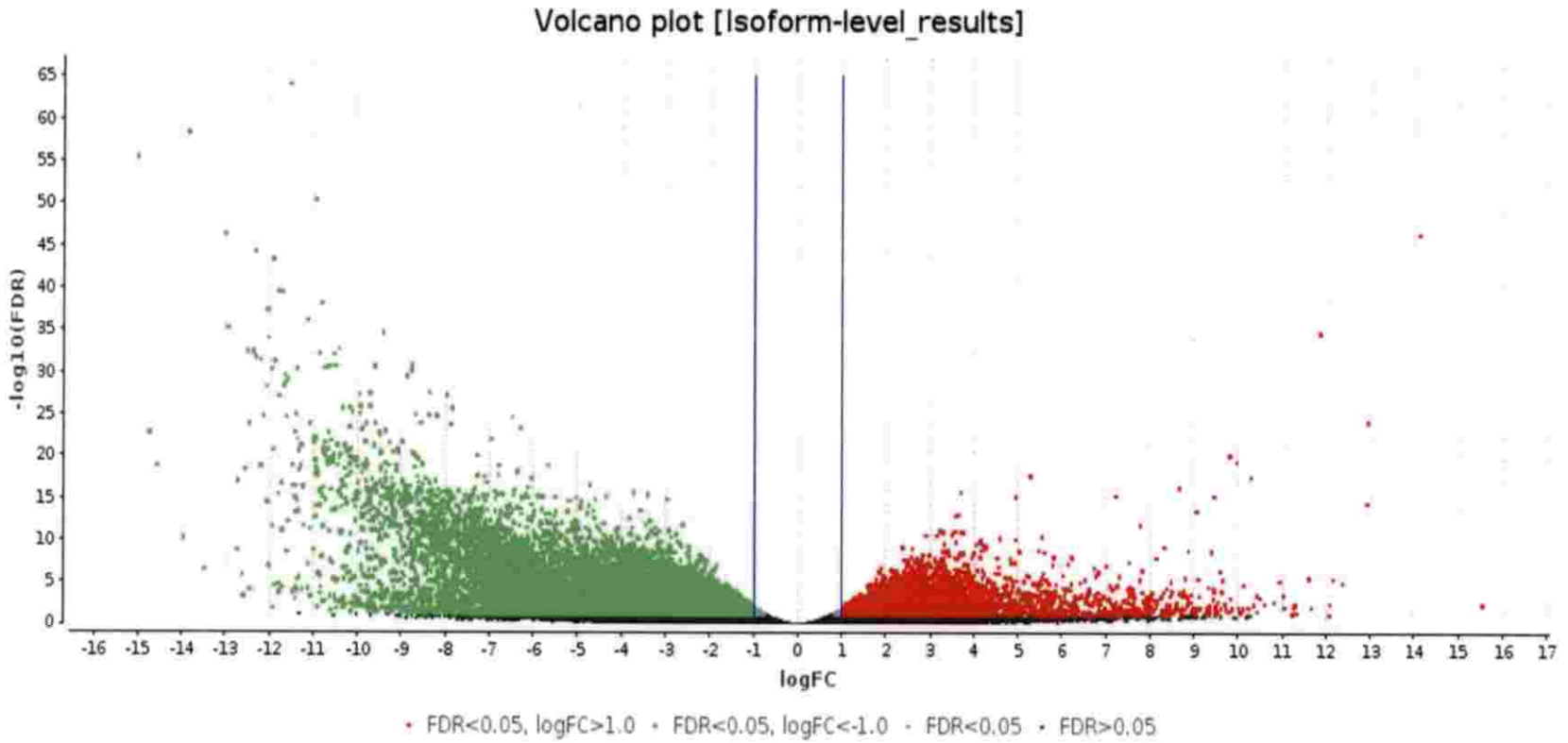


Figure 15. Heatmap showing the top 50 differentially expressed genes betweenpurple and white pair analysis

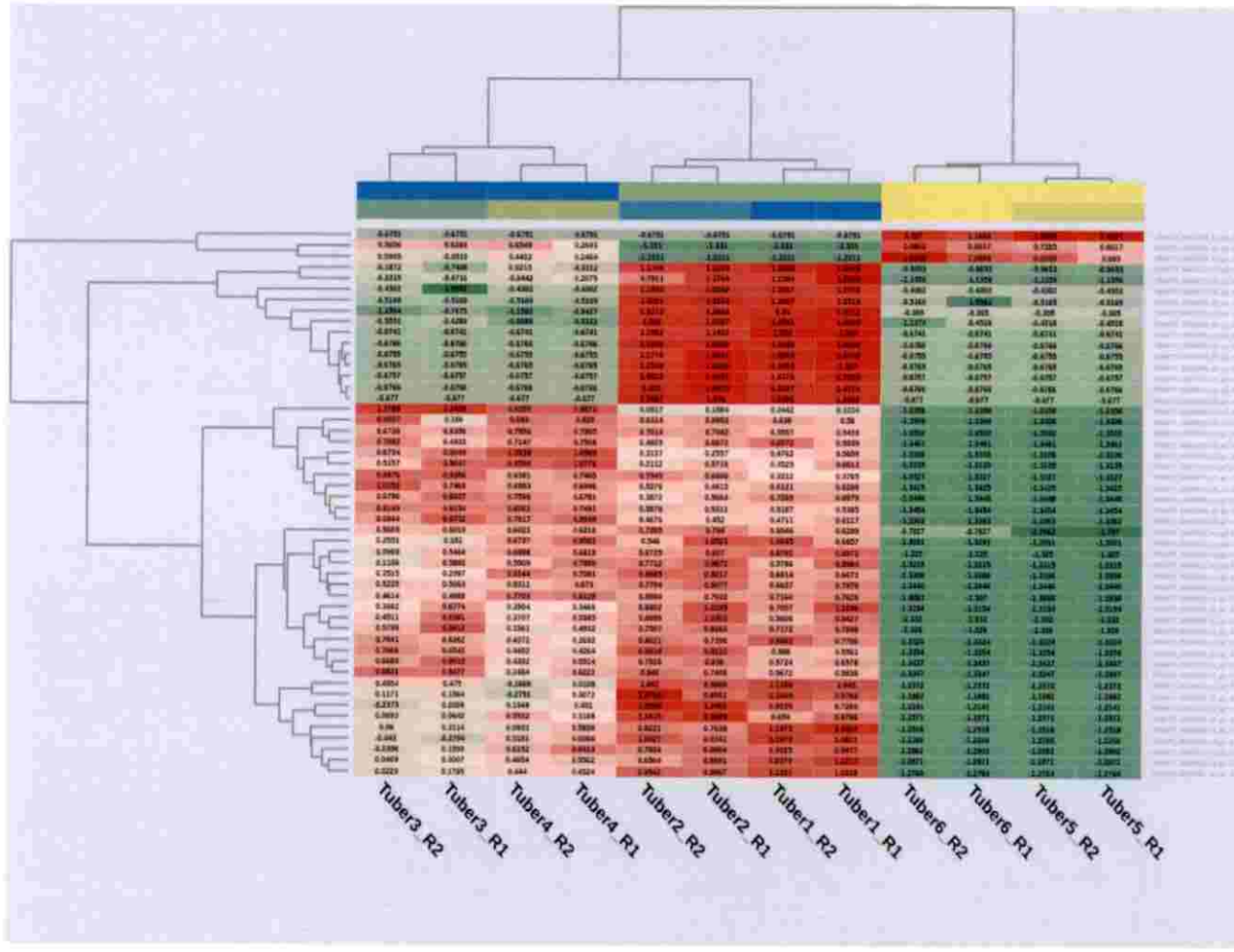


Figure 16. Venn diagram of all upregulated DEGs between the three pairwise analysis libraries

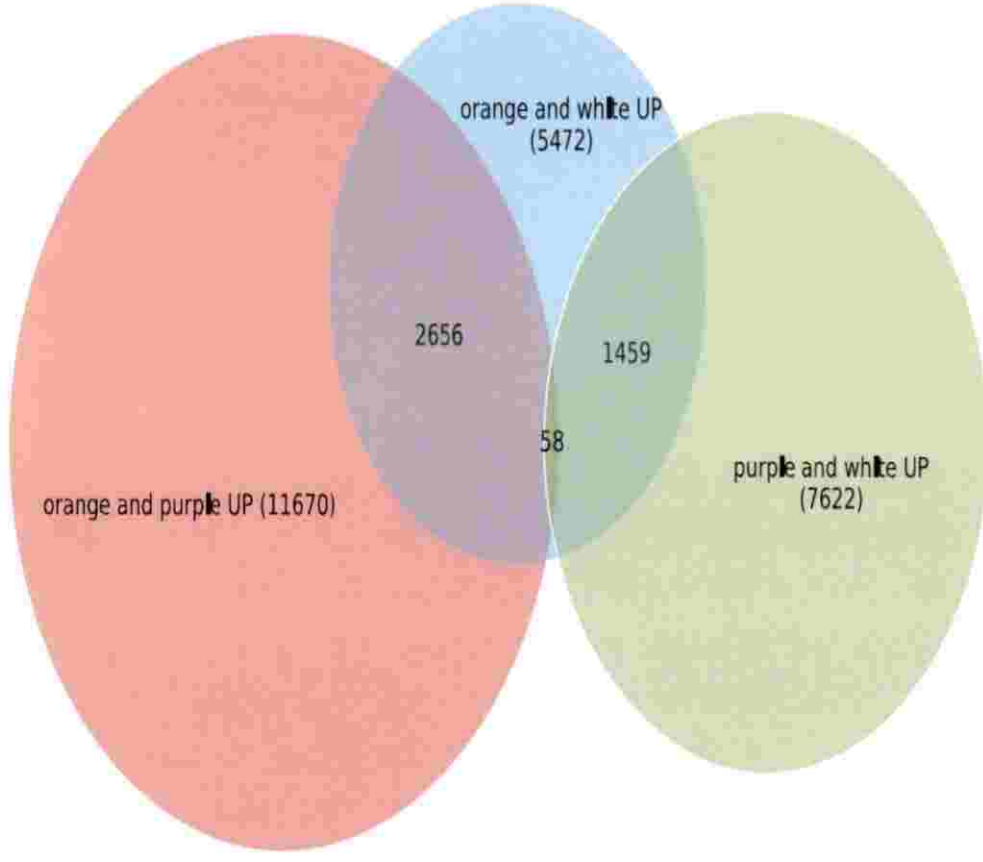
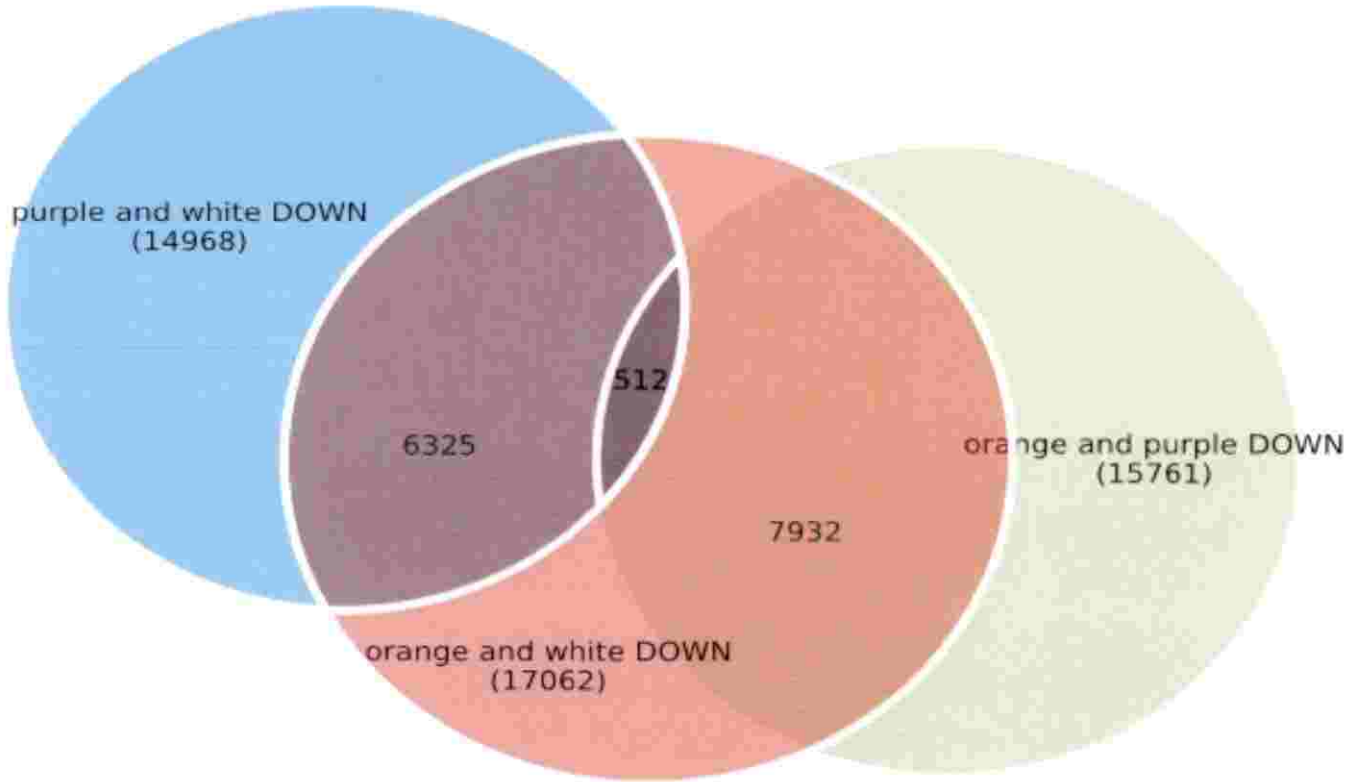


Figure 17. Venn diagram of all downregulated DEGs between the three pairwise analysis libraries



4.1.6. Enrichment analysis of Differentially Expressed Genes

The enrichment analysis of the differentially expressed genes were done using both Fischer's exact test and GSEA method. Both the method searches for pathways whose genes are significantly enriched (i.e. over-represented) in the fixed list of genes of interest, compared to all genes in the *de novo* assembled transcriptome. GSEA is a threshold-free method that analyzes all genes based on their differential expression rank or other score, without prior gene filtering. Upon enrichment of differentially expressed genes obtained from orange and white pairwise analysis, the upregulated genes were significantly enriched for cell periphery, organonitrogen compound biosynthesis process, plasma membrane (Figure18). The downregulated genes were significantly enriched for regulation of metabolic process, lipid metabolic process, cell wall macromolecule metabolic process (Figure 19). The GSEA enrichment of differentially expressed genes generated the top core enriched gene ontology term for each annotation includes amide transport, peptide transport, protein localization (Figure 20). The enrichment analysis of differentially expressed genes found from orange and purple pairwise analysis, the upregulated genes showed the maximum percentage of sequences for the gene ontology terms catalase activity, membrane and nitrogen compound metabolic process (Figure21). The core enriched gene ontology term from GSEA analysis were found to be cellular lipid metabolic process, ribonucleoprotein complex, transferase activity (Figure 22).The differentially expressed genes obtained from purple and white pairwise analysis upon enrichment showed the maximum percentage of sequences for the gene ontology terms membrane, membrane part, intrinsic component of membrane for the upregulated genes category (Figure 23) and catalytic activity, membrane and nitrogen compound metabolic process for downregulated genes category (Figure 24). The core enriched GO terms from GSEA were found to be amide transport, protein localization and the peptide transport (Figure 25).

Figure 18. Bar chart showing the percentages of sequences for each annotation for upregulated genes between orange and white pairwise analysis

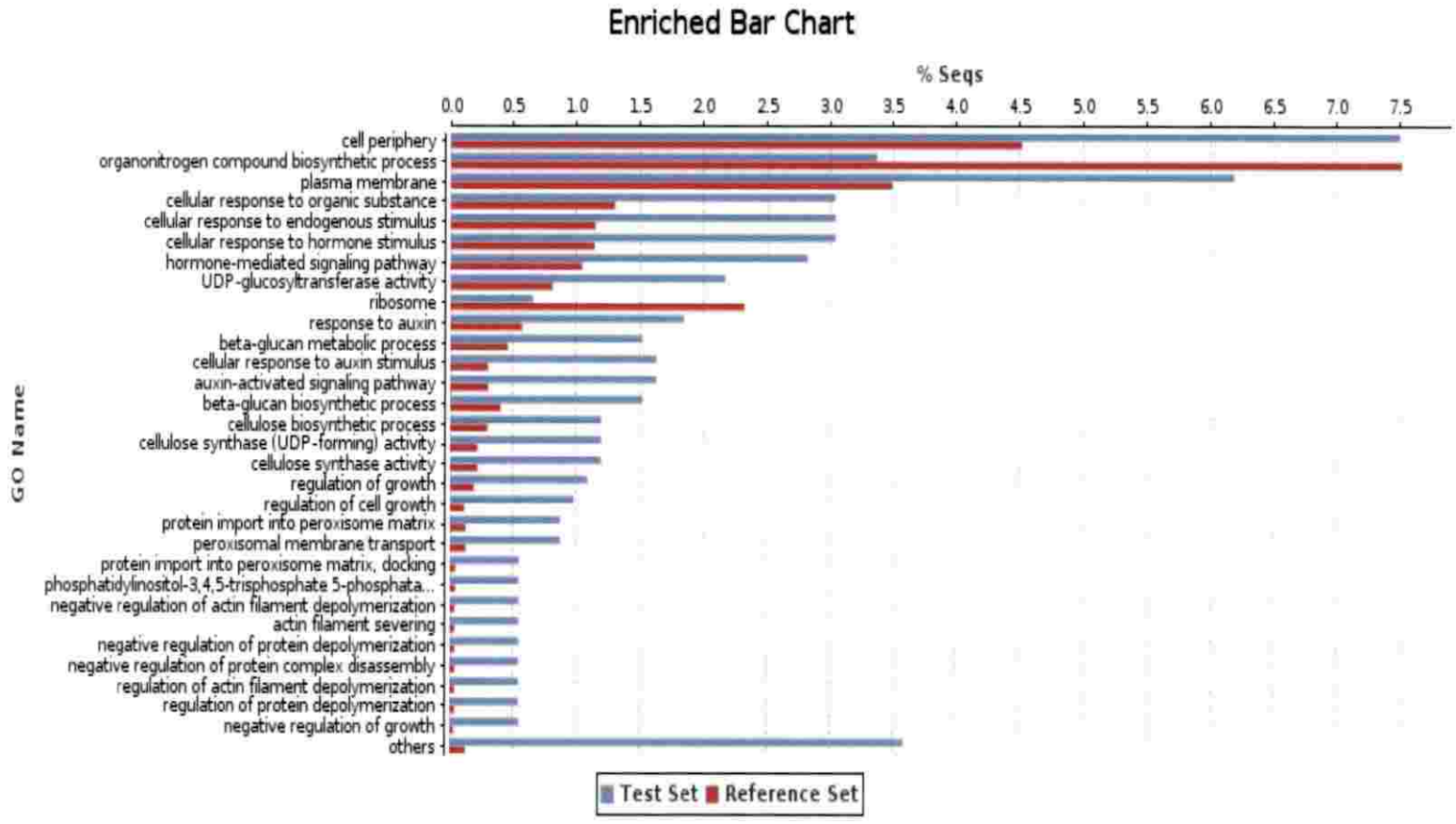


Figure 19. Bar chart showing the percentages of sequences for each annotation for downregulated genes between orange and white pairwise analysis

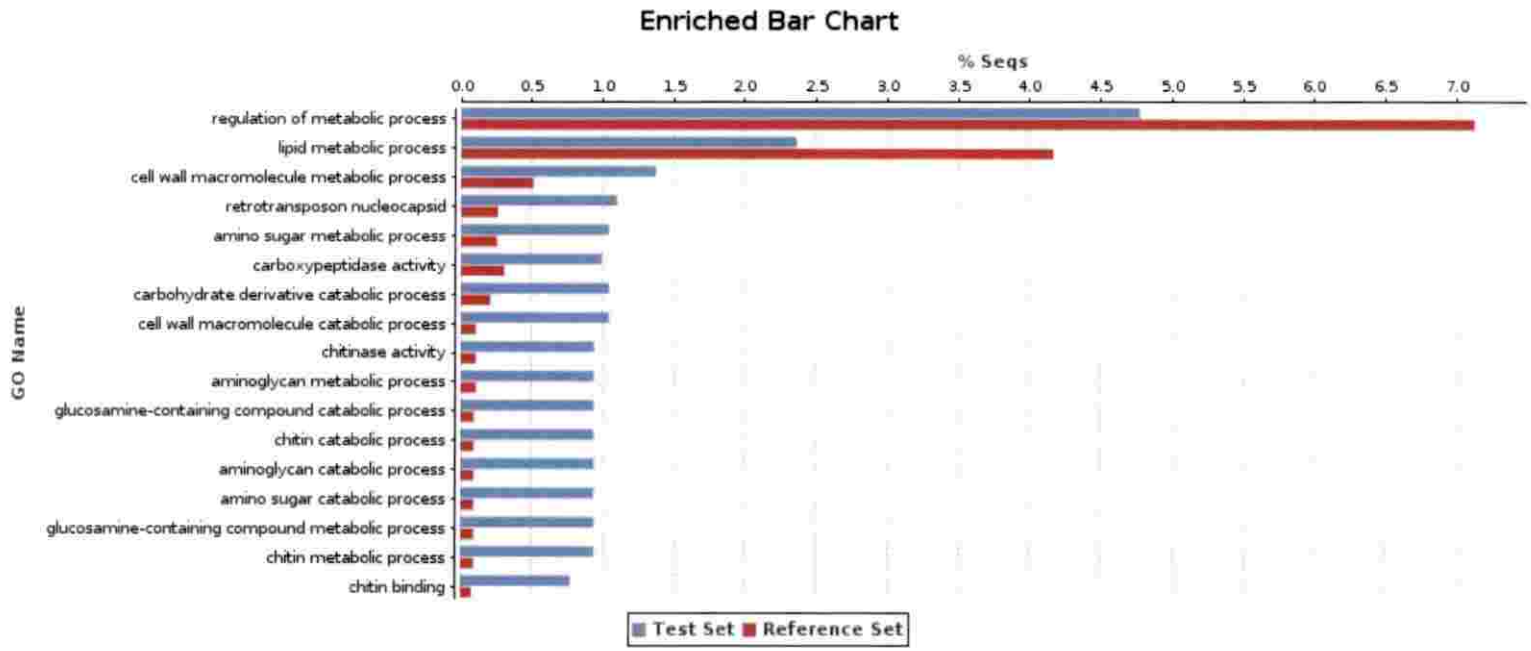


Figure 20. Bar chart showing core enriched GO for each annotation obtained for orange and white pairwise analysis

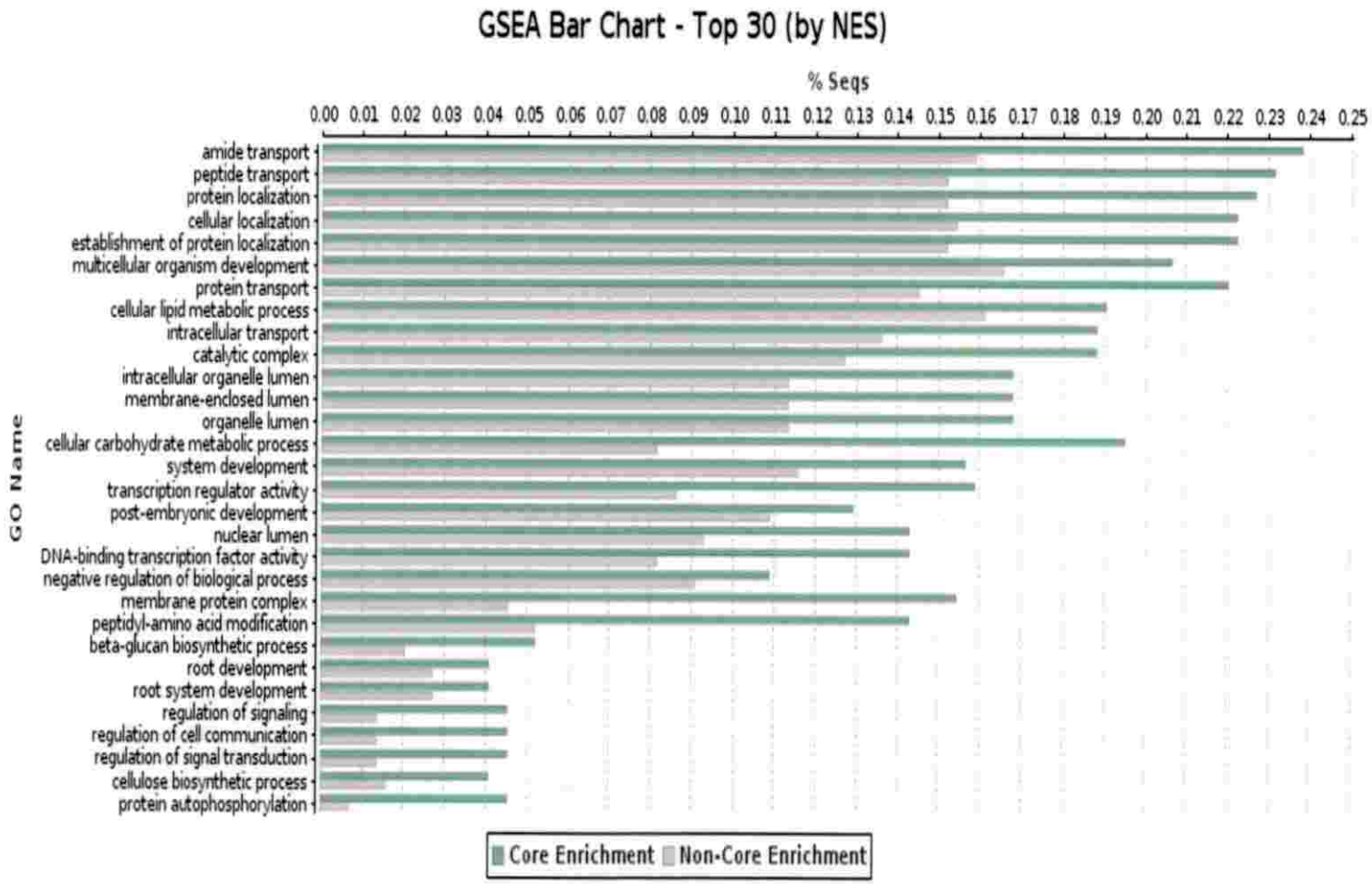


Figure 21. Bar chart showing the percentages of sequences for each annotation for upregulated genes between orange and purple pairwise analysis

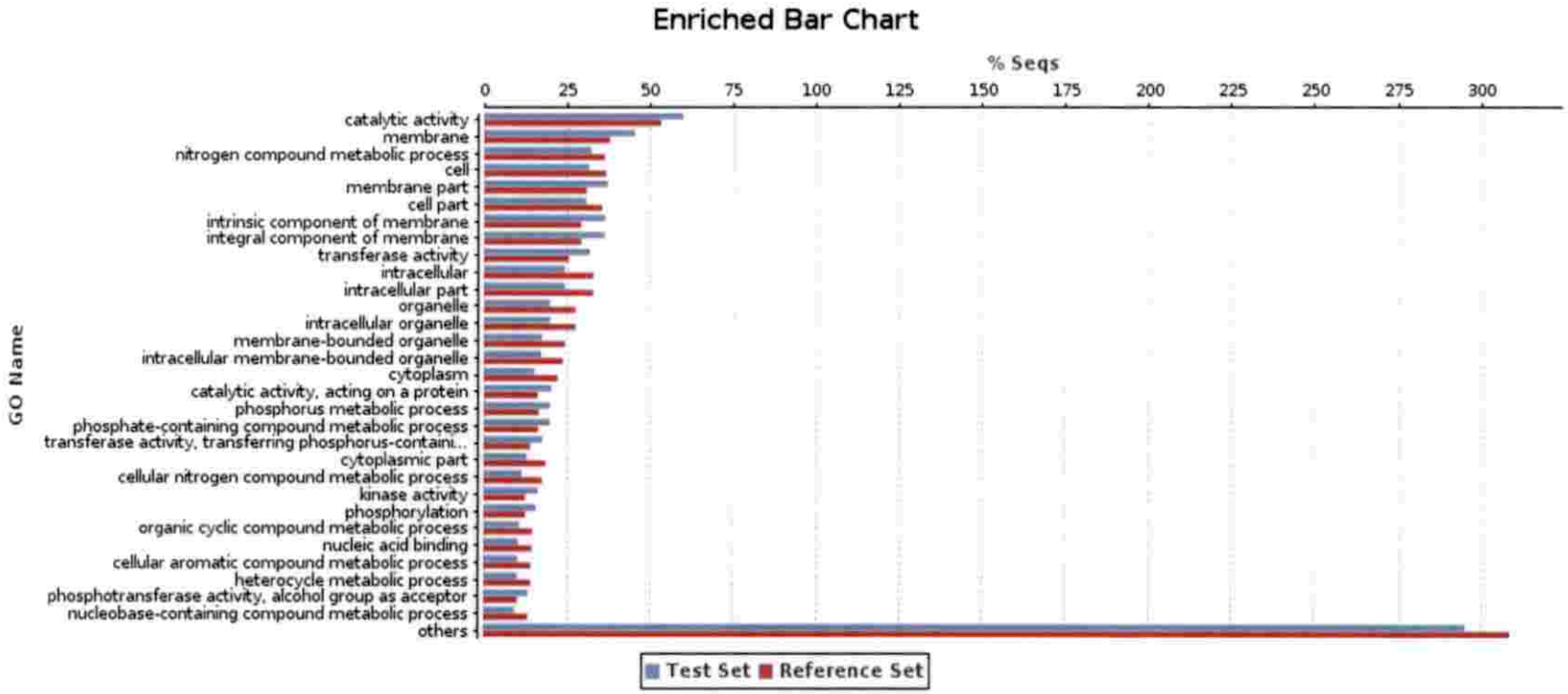


Figure 22. Bar chart showing core enriched GO for each annotation obtained for orange and purple pairwise analysis

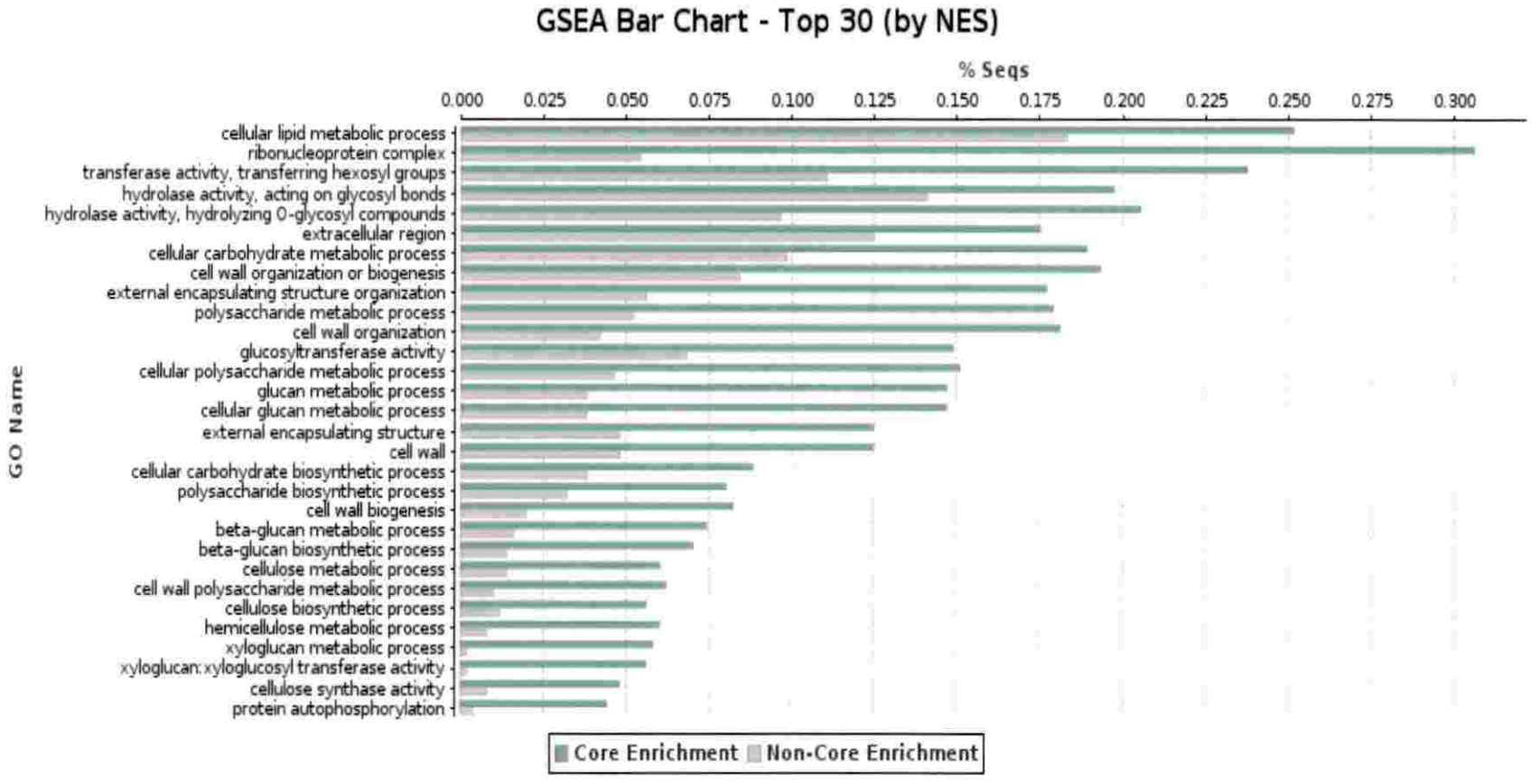


Figure 23. Bar chart showing the percentages of sequences for each annotation for upregulated genes between purple and white pairwise analysis

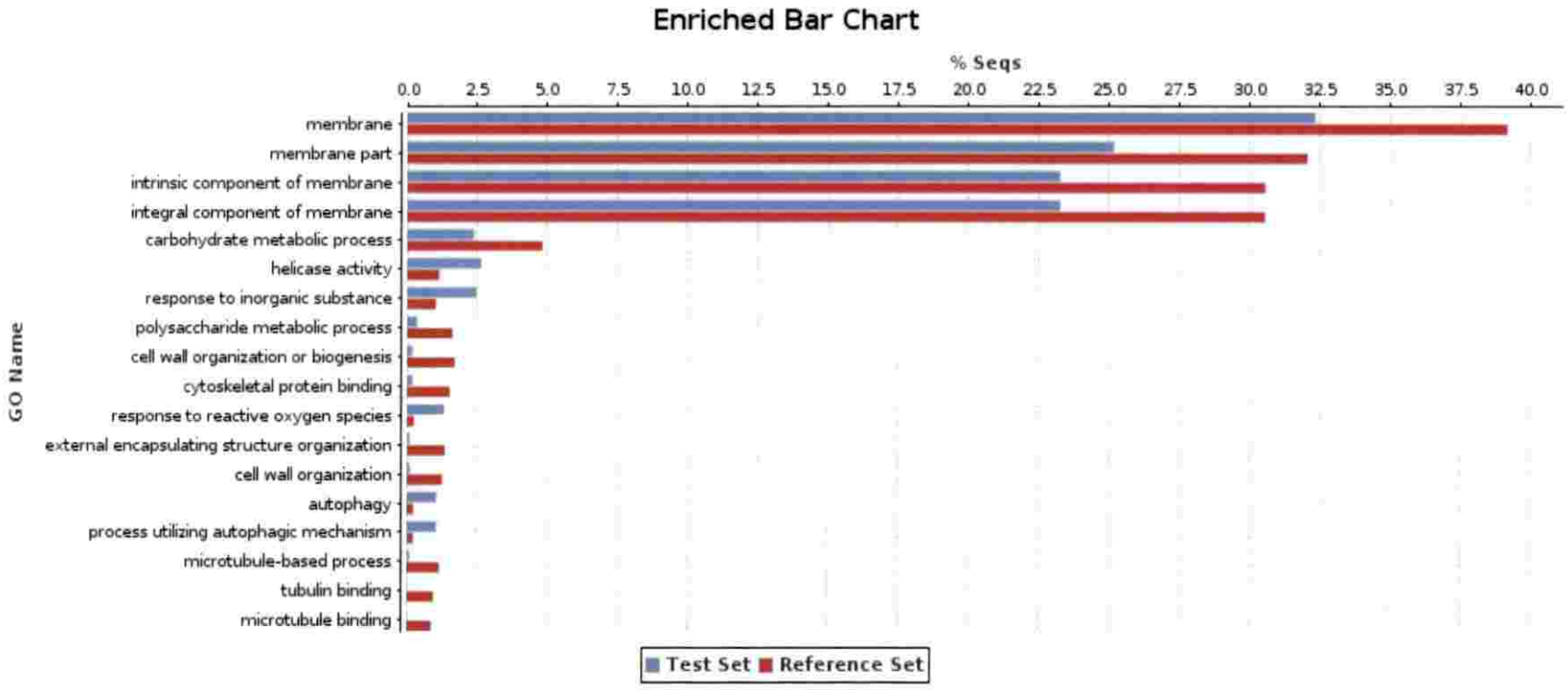


Figure 24. Bar chart showing the percentages of sequences for each annotation of downregulated genes in purple and white pairwise analysis

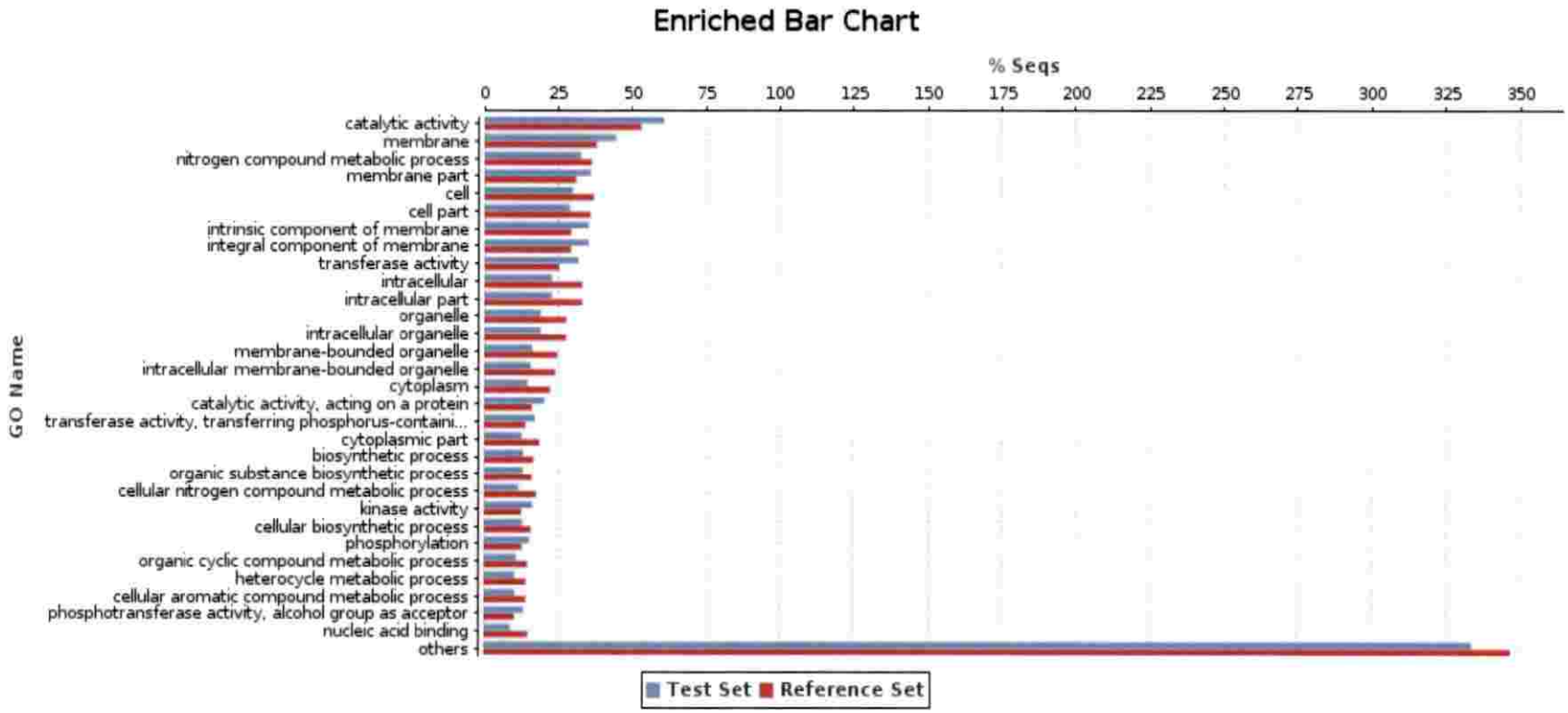
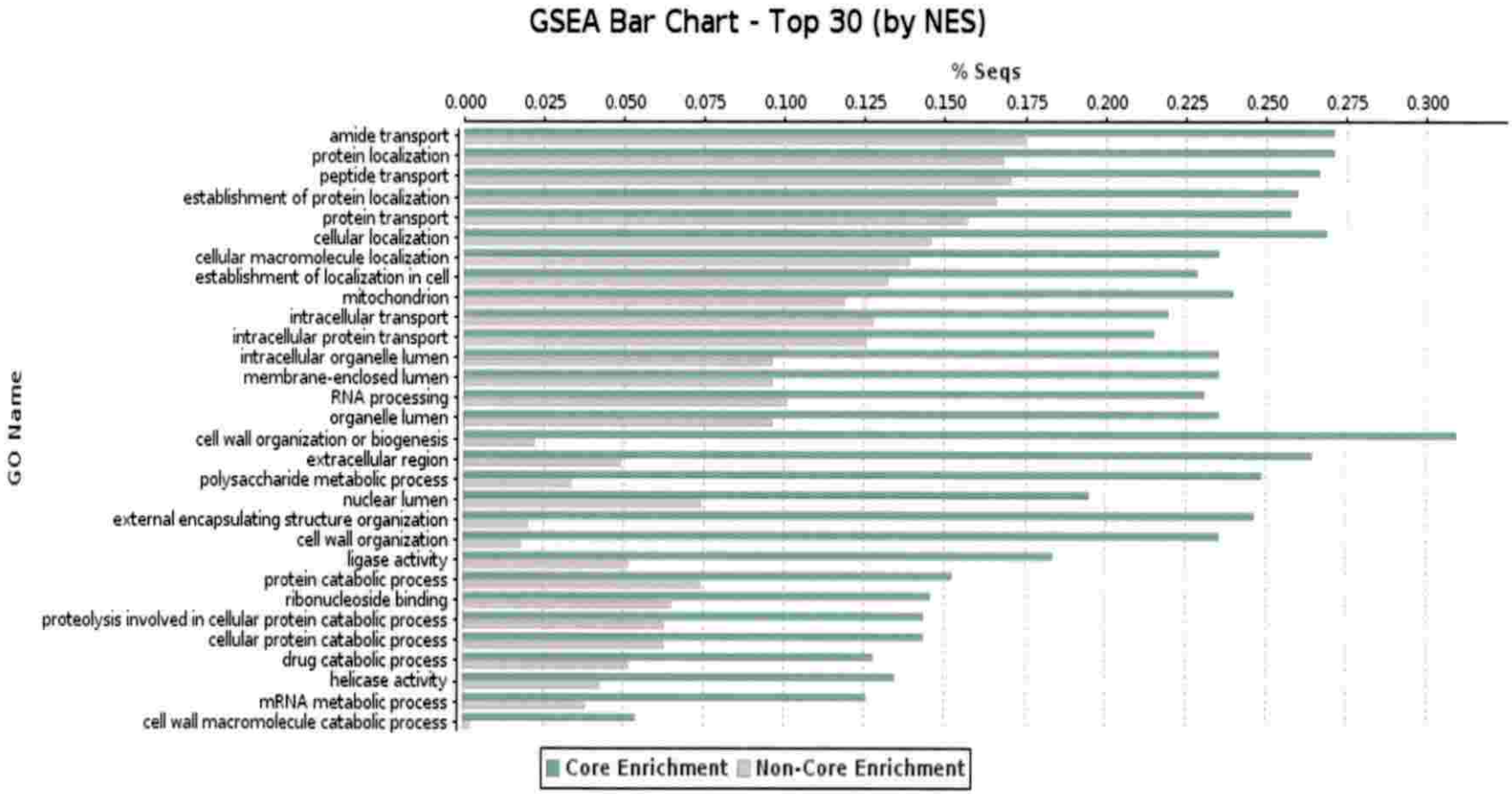


Figure 25. Bar chart showing core enriched GO for each annotation obtained for purple and white pairwise analysis



4.1.7. Functional annotation of core enriched terms

The core enriched terms from GSEA for each of the three pairwise analysis was searched for the terms related to the pigment biosynthesis pathway. As a result, the gene ontology terms Isoprenoid biosynthesis and metabolic process terms were found to be enriched for the orange and white pairwise analysis, terpenoid metabolic process for orange and purple pairwise analysis and antioxidant activity for the purple and white pairwise analysis. The transcripts associated with these enriched terms were obtained with their enrichment rank and the transcript sequences were retrieved from the *de novo* assembled transcriptome assembly. Thus obtained transcript sequences for each of the core enriched terms related to pigment biosynthesis pathway for all the three pairwise analysis was functionally annotated to identify the differentially expressed genes. For the enriched term isoprenoid biosynthetic process, twenty-four transcripts were found to be associated with the enriched term depending on the enrichment score (Table14). Upon functional annotation of the twenty-four transcripts sequences, genes involved in the carotenoid biosynthesis pathway including β-carotene hydroxylase, zeta-carotene desaturase, isopentenyl-diphosphate (IPI), squalene synthase, geranyl-geranyl diphosphate synthase, zeaxanthin epoxidase were obtained. The two significant alignments of pathway genes transcript sequences are shown (Table15). The core enriched term isoprenoid metabolic process obtained from orange and white pairwise analysis, produced an account of twenty-nine transcripts associated with the enriched term (Table 16). Functional annotation of transcript sequences associated with isoprenoid metabolic process retrieved from *de novo* assembled transcriptome resulted in the identification of zeaxanthin epoxidase, phytoene synthase, β-carotene hydroxylase, isopentenyl-diphosphate (IPI), squalene synthase, geranyl-geranyl diphosphate synthase and zeta-carotene desaturase involved in carotenoid biosynthesis pathway and the two significant alignments of these identified genes are shown (Table 17). The core enriched term related to pigment biosynthesis pathway of orange and purple pairwise analysis was terpenoid metabolic process. Thirteen transcripts related to terpenoid metabolic process were functionally annotated after retrieving transcript

sequences from *de novo* assembled transcriptome assembly (Table 18). β-carotene hydroxylase, geranyl-geranyl diphosphate synthase and zeta-carotene desaturase genes involved in carotenoid biosynthesis pathway were identified (Table 19). Purple and white pairwise analysis didn't show any enriched terms related to pigment biosynthesis. The term antioxidant activity was included in the core enriched term was chosen as the colour imparting pigment in purple fleshed sweet potato is anthocyanin which is the major component for the antioxidant activity. Fourteen transcripts related to antioxidant activity were obtained (Table20) and the sequences retrieved from the *de novo* assembled transcriptome were functionally annotated to identify the significant alignments (Table21).

Table 14. List of Core Enriched Sequences for the term isoprenoid biosynthetic process

Probe Rank in Gene List	Rank	Core Enrichment
TRINITY_DN15343_c0_g1_i1	1675	Yes
TRINITY_DN22722_c1_g1_i3	4304	Yes
TRINITY_DN1998_c0_g1_i1	4502	Yes
TRINITY_DN24255_c0_g1_i2	5852	Yes
TRINITY_DN12813_c0_g1_i1	7363	Yes
TRINITY_DN22722_c1_g1_i2	7957	Yes
TRINITY_DN18631_c0_g1_i4	8962	Yes
TRINITY_DN27548_c1_g1_i5	17133	Yes
TRINITY_DN28226_c0_g2_i2	18548	Yes
TRINITY_DN22722_c1_g1_i1	10790	Yes
TRINITY_DN25828_c0_g1_i1	12107	Yes
TRINITY_DN22644_c0_g1_i2	14443	Yes

Table. 14 continued

TRINITY_DN24053_c1_g1_i8	14663	Yes
TRINITY_DN6766_c0_g1_i1	14807	Yes
TRINITY_DN29410_c2_g1_i2	19241	Yes
TRINITY_DN27548_c1_g1_i6	20607	Yes
TRINITY_DN16357_c0_g1_i3	21621	Yes
TRINITY_DN29410_c2_g1_i1	23249	Yes
TRINITY_DN16357_c0_g1_i1	23499	Yes
TRINITY_DN41320_c0_g1_i1	24245	Yes
TRINITY_DN22791_c0_g1_i1	26202	Yes
TRINITY_DN32392_c0_g1_i1	26831	Yes
TRINITY_DN22791_c0_g1_i2	29113	Yes
TRINITY_DN29410_c2_g1_i4	29776	Yes

Table 15. List of differentially expressed genes for the enriched term isoprenoid biosynthetic process

Query Name	Significant Alignments	Scientific Taxonomy	E-Value	Sim.	Acc.	Gene Name (Taxa Id)	Xref (DB)	GO. Mapping
TRINITY_DN18631_c0_g1_i1	**PREDICTED: beta-carotene hydroxylase 2, chloroplastic**	*Ipomoea nil*	4.24914e-113	100.0%	XP_019195604	-	-	-
TRINITY_DN18631_c0_g1_i1	**beta-carotene hydroxylase**	*Ipomoea nil*	4.29587e-113	100.0%	BAI47580	CHYB(35883)	BAI47580 (InterPro)	GO:0005506 GO:0055114 GO:0016021 GO:0016020 GO:0008610 GO:0016491
TRINITY_DN28226_c0_g2_i2	**PREDICTED: Zeta-carotene desaturase, chloroplastic/chromoplastic**	*Ipomoea nil*	8.619287e-120	98.9%	XP_019187856			
TRINITY_DN28226_c0_g2_i2	**Zeta-carotene desaturase, partial**	*Vaccinium myrtillus*	2.4749e-108	91.3%	ANC48851	ZDS(180763)	ANC4881 (InterPro)	GO:0016491 GO:0055114
TRINITY_DN22644_c0_g1_i2	**Isopentenyl diphosphate isomerase**	*Ipomoea batatas*	6.04078e-123	100.0%	AAZ94730	ipi(4120)	AAZ9470 (InterPro)	GO:0004452 GO:0016853 GO:0016787 GO:0008299

TRINITY_DN22644_c0_g1_i2	**PREDICTED: isopentenyl-diphosphate Delta-isomerase I-like**	*Ipomoea nil*	5.7802 8e-121	99.4%	XP_019194 619	-	-	-
TRINITY_DN22722_c1_g1_i1	**PREDICTED:Squalene synthase**	*Ipomoea nil*	0	99%	XP_019169 346 XP_019169 347 XP_019169 348	-	-	-
TRINITY_DN22722_c1_g1_i1	**Squalene synthase**	*Capsicum chinense*	0	95.4%	PHU10676	BC332_22536 (80379)	PHU1066 (InterPro)	GO:0051996 GO:0009058 GO:0016020 GO:0008610 GO:0004310 GO:0006696 GO:0016740 GO:0016021 GO:0016765
TRINITY_DN24255_c0_g1_i2	**Geranylgeranyl diphosphate synthase**	*Ipomoea batatas*	0	99.7%	ACF37217	G8XQT1(4120)	ACF37217 (InterPro)	GO:0008299 GO:0016740
TRINITY_DN24255_c0_g1_i2	**Prenyltransferase**	*Ipomoea batatas*	0	99.4%	AIP89952	-	-	-

TRINITY_DN22722_c1_g1_i2	**PREDICTED: Squalene synthase**	*Ipomoea nil*	0	99.9%	XP_019169346 XP_019169347 XP_019169348	-	-	-
TRINITY_DN22722_c1_g1_i2	**Squalene synthase**	*Capsicum chinense*	0	95.4%	PHU10676	BC332_22536 (80379)	PHU1066 (InterPro)	GO:0051996 GO:0009058 GO:0016020 GO:0008610 GO:0004310 GO:0006696 GO:0016740 GO:0016021 GO:0016765
TRINITY_DN22722_c1_g1_i3	**PREDICTED:Squalene synthase**	*Ipomoea nil*	0	98.9%	XP_019169346 XP_019169347 XP_019169348	-	-	-

TRINITY_DN22722_c1_g1_i3	**Squalene synthase**	*Capsicum chinense*	0	95.7	PHU10676	BC332_22536 (80379)	PHU1066 (InterPro)	GO:0051996 GO:0009058 GO:0016020 GO:0008610 GO:0004310 GO:0006696 GO:0016021 GO:0016765
TRINITY_DN1998_c0_g1_i1	**Zeaxanthin epoxidase, partial**	*Ipomoea batatas*	1.9319 3e-42	99.6%	AIR95905	A0A1L1WGL6 (4120)	AIR95905 (InterPro)	GO:0009507 GO:0009540 GO:0016020 GO:0009688 GO:0055114
TRINITY_DN1998_c0_g1_i1	**PREDICTED: Zeaxanthin epoxidase, chloroplastic-like**	*Ipomoea nil*	3.5891 5e-41	96.0%	XP_019195 269	-	-	-

Table 16. List of Core Enriched Sequences for the term isoprenoid metabolic process

Probe	Rank in Gene List	Rank Metric	Core Enrichment
TRINITY_DN15343_c0_g1_i1		1675	Yes
TRINITY_DN27079_c0_g1_i3		2295	Yes
TRINITY_DN22722_c1_g1_i3		4304	Yes
TRINITY_DN1998_c0_g1_i1		4502	Yes
TRINITY_DN24255_c0_g1_i2		5852	Yes
TRINITY_DN12813_c0_g1_i1		7363	Yes
TRINITY_DN22722_c1_g1_i2		7957	Yes
TRINITY_DN18631_c0_g1_i4		8962	Yes
TRINITY_DN22722_c1_g1_i1		10790	Yes
TRINITY_DN29164_c0_g1_i4		10828	Yes
TRINITY_DN25828_c0_g1_i1		12107	Yes
TRINITY_DN22644_c0_g1_i2		14443	Yes
TRINITY_DN24053_c1_g1_i8		14663	Yes

Table .16 contd.

TRINITY_DN6766_c0_g1_i1	14807	Yes
TRINITY_DN27548_c1_g1_i5	17133	Yes
TRINITY_DN27090_c0_g1_i1	17722	Yes
TRINITY_DN28226_c0_g2_i2	18548	Yes
TRINITY_DN29410_c2_g1_i2	19241	Yes
TRINITY_DN27548_c1_g1_i6	20607	Yes
TRINITY_DN27041_c0_g1_i2	21038	Yes
TRINITY_DN16357_c0_g1_i3	21621	Yes
TRINITY_DN29410_c2_g1_i1	23249	Yes
TRINITY_DN16357_c0_g1_i1	23499	Yes
TRINITY_DN41320_c0_g1_i1	24245	Yes
TRINITY_DN22791_c0_g1_i1	26202	Yes
TRINITY_DN32392_c0_g1_i1	.26831	Yes

Table. 16 contd

.

TRINITY_DN29164_c1_g3_i5	27112	Yes
TRINITY_DN22791_c0_g1_i2	29113	Yes
TRINITY_DN29410_c2_g1_i4	29776	Yes

Table 17. List of differentially expressed genes for the enriched term isoprenoid metabolic process

Query Name	Significant Alignments	Scientific Taxonomy	E-Value	Sim.	Acc.	Gene Name (Taxa Id)	Xref (DB)	G.O. Mapping
TRINITY_DN1998_c0_g1_i1	**Zeaxanthin epoxidase, partial**	*Ipomoea batatas*	1.93193e-42	100 %	AIR95905	A0A1L1WGL6 (4120)	AIR95905 (InterPro)	GO:0071949 GO:0009507 GO:0009540 GO:0016020 GO:0009688 GO:0055114
TRINITY_DN1998_c0_g1_i1	**PREDICTED: Zeaxanthin epoxidase, chloroplastic-like**	*Ipomoea nil*	3.58915e-41	96%	XP_019195269	-	-	-
TRINITY_DN15343_c0_g1_i1	**PREDICTED: Phytoene synthase 2, chloroplastic-like**	*Ipomoea nil*	0	97.8 %	XP_019182942	-	-	-
TRINITY_DN15343_c0_g1_i1	**PREDICTED: Phytoene synthase PREDICTED: phytoene synthase 2, chloroplasticstic**	*Nicotiana sylvestris Nicotiana tabacum*	0	90%	XP_009770434 XP_009770435 XP_016435460 XP_016435461	LOC107761716 (4097) LOC104221141 (4096)	XP_009770434 (InterPro) XP_016435460 (InterPro)	GO:0051996 GO:0006696 GO:0016765 GO:0004310

TRINITY_DN18631_c0_g1_i4	**beta-ring hydroxylase, partial**	*Gardenia jasminoides*	9.98984e-36	100 %	AFI09272	I1U6J2 (114476)	AFI09272 (InterPro)	GO:0016020 GO:0016021 GO:0016491 GO:0005506 GO:0055114 GO:0008610
TRINITY_DN18631_c0_g1_i4	**beta-carotene hydroxylase**	*Ipomoea sp. Kenyan*	1.03081e-34	100 %	BAI47578	CHYB(641442)	BAI47578 (InterPro)	GO:0055114 GO:0005506 GO:0008610 GO:0016491 GO:0016021 GO:0016020
TRINITY_DN22644_c0_g1_i2	**Isopentenyl diphosphate isomerase**	*Ipomoea batatas*	6.04078e-123	100 %	AAZ94730	ipi(4120)	AAZ94730 (InterPro)	GO:0004452 GO:0016853 GO:0016787 GO:0008299
TRINITY_DN22644_c0_g1_i2	**PREDICTED: Isopentenyl-diphosphate Delta-isomerase I-like**	*Ipomoea nil*	5.78028e-121	99.4 %	XP_019194619	-	-	-
TRINITY_DN22722_c1_g1_i2	**PREDICTED: Squalene synthase**	*Ipomoea nil*	0	99%	XP_019169346 XP_019169347 XP_019169348	-	-	-

TRINITY_DN22722_c1_g1_i2	**Squalene synthase**	*Capsicum chinense*	0	95.4 %	PHU10676	BC332_22536 (80379)	PHU10676 (InterPro)	GO:0051996 GO:0009058 GO:0016020 GO:0008610 GO:0004310 GO:0006696 GO:0016740 GO:0016021 GO:0016765
TRINITY_DN24255_c0_g1_i2	**Geranylgeranyl diphosphate synthase**	*Ipomoea batatas*	0	99.7 %	ACF37217	G8XQT1(4120)	ACF37217 (InterPro)	GO:0008299 GO:0016740
TRINITY_DN24255_c0_g1_i2	**Prenyltransferase**	*Ipomoea batatas*	0	99.4 %	AIP89952	-	-	-
TRINITY_DN28226_c0_g1_i2	**PREDICTED: Zeta-carotene desaturase, chloroplastic/chromoplastic**	*Ipomoea nil*	8.61927e-120	98.9 %	XP_019187856	-	-	-
TRINITY_DN28226_c0_g1_i2	**Zeta-carotene desaturase, partial**	*Vaccinium myrtillus*	2.4749e-108	91.3 %	ANC48851	ZDS(180763)	ANC48851 (InterPro)	GO:0016491 GO:0055114

Table 18. List of Core Enriched Sequences for the term terpenoid metabolic process

Probe Rank in Gene List	Rank Metric	Core Enrichment
TRINITY_DN29410_c2_g1_i1	2423	Yes
TRINITY_DN33455_c0_g1_i1	3005	Yes
TRINITY_DN29410_c2_g1_i2	3740	Yes
TRINITY_DN32891_c0_g1_i1	4132	Yes
TRINITY_DN29164_c0_g1_i4	10214	Yes
TRINITY_DN29164_c1_g3_i4	10744	Yes
TRINITY_DN29410_c2_g1_i4	11153	Yes
TRINITY_DN18631_c0_g1_i4	11299	Yes
TRINITY_DN24255_c0_g1_i2	11421	Yes
TRINITY_DN29164_c1_g3_i2	14211	Yes
TRINITY_DN27079_c0_g1_i2	14908	Yes
TRINITY_DN20726_c0_g1_i1	17227	Yes
TRINITY_DN28226_c0_g2_i3	17401	Yes

Table 19. List of differentially expressed genes for the enriched term terpenoidmetabolic process

Query Name	Significant Alignments	Scientific Taxonomy	E-Value	Sim.	Acc.	Gene Name (Taxa Id)	Xref (DB)	Mapping
TRINITY_DN18631_c0_g1_i4	**beta-ring hydroxylase, partial**	*Gardenia jasminoides*	9.9898 4e-36	100%	AFI09272	I1U6J2 (114476)	AFI09272 (InterPro)	GO:0016020 GO:0016021 GO:0016491 GO:0005506 GO:0055114 GO:0008610
TRINITY_DN18631_c0_g1_i4	**beta-carotene hydroxylase**	*Ipomoea sp. Kenyan*	1.0308 1e-34	100%	BAI47578	CHYB (641442)	BAI47578 (InterPro)	GO:0055114 GO:0005506 GO:0008610 GO:0016491 GO:0016021 GO:0016020
TRINITY_DN24255_c0_g1_i2	**Geranylgeranyl diphosphate synthase**	*Ipomoea batatas*	0	99.7%	ACF37217	G8XQT1 (4120)	ACF3727 (InterPro)	GO:0008299 GO:0016740

TRINITY_DN24255_c0_g1_i2	**Prenyltransferase**	*Ipomoea batatas*	0	99.4%	AIP89952	-	-	-
TRINITY_DN28226_c0_g2_i3	**PREDICTED: Zeta-carotene desaturase, chloroplastic/chromoplastic**	*Ipomoea nil*	3.40245e-67	99.1%	XP_019187856	-	-	-
TRINITY_DN28226_c0_g2_i3	**Zeta-carotene desaturase, chloroplastic/chromoplastic-like**	*Citrus sinensis*	.15413e-58	98%	XP_006471845	-	-	-

Table 19 contd.

Table 20. List of Core Enriched Sequences for the term antioxidant activity

Probe Rank in Gene List	Rank Metric	Core Enrichment
TRINITY_DN26060_c0_g1_i6	25836	Yes
TRINITY_DN33946_c0_g1_i1	25981	Yes
TRINITY_DN13064_c0_g1_i1	26006	Yes
TRINITY_DN12242_c0_g1_i1	26464	Yes
TRINITY_DN23847_c0_g1_i1	27088	Yes
TRINITY_DN12205_c0_g1_i1	28761	Yes
TRINITY_DN38415_c0_g1_i1	29680	Yes
TRINITY_DN29131_c0_g1_i4	30065	Yes
TRINITY_DN26060_c0_g1_i4	30297	Yes
TRINITY_DN20238_c0_g1_i4	32150	Yes
TRINITY_DN27562_c0_g1_i1	33032	Yes
TRINITY_DN19015_c0_g1_i1	33958	Yes
TRINITY_DN27562_c1_g1_i3	37908	Yes

Table 21. List of differentially expressed genes for the enriched term antioxidant activity process

Query Name	Significant Alignments	Scientific Taxonomy	E-Value	Sim.	Acc.	Gene Name (Taxa Id)	Xref (DB)	G.O. Mapping
TRINITY_DN13064_c0_g1_i1	**PREDICTED: peroxidase 11-like**	*Ipomoea nil*	2.63255e-82	96.3%	XP_019149837	-	-	-
TRINITY_DN13064_c0_g1_i1	**PREDICTED: peroxidase 11-like**	*Nicotiana sylvestris*	1.64895e-52	80.1%	XP_009773752	LOC104223917 (4096)	XP_009773752 (InterPro)	GO:0016491 GO:0055114 GO:0046872 GO:0004601 GO:0006979 GO:0042744 GO:009886 GO:0020037 GO:0005576
TRINITY_DN19015_c0_g1_i1	**PREDICTED: 1-Cys peroxiredoxin A**	*Ipomoea nil*	6.07144e-155	98.6%	XP_019193381	-	-	-
TRINITY_DN19015_c0_g1_i1	**PREDICTED: probable 1-Cys peroxiredoxin**	*Ipomoea nil*	6.32848e-136	93.9%	XP_019176086	-	-	-

TRINITY_DN23847_c0_g1_i1	**PREDICTED: peroxidase 17**	*Daucus carota subsp. sativus*	5.55057e-157	93.7%	XP_017229533 KZN10820	DCAR_003476 (79200)	XP_017229533 (InterPro)	GO:0055114 GO:0098869 GO:0042744 GO:0046872 GO:0016491 GO:0004601 GO:0020037 GO:0005576 GO:0006979
TRINITY_DN23847_c0_g1_i1	**Peroxidase 17-like**	*Camellia sinensis*	5.5753e-157	93.2%	XP_028067738	-	-	-
TRINITY_DN27562_c0_g1_i1	**basic peroxidase swpb4**	*Ipomoea batatas*	6.70276e-74	99.3%	ABR23054	-	-	-
TRINITY_DN27562_c0_g1_i1	**PREDICTED: peroxidase P7**	*Ipomoea nil*	1.82906e-72	97.8%	XP_019200127	-	-	-

Table 21 contd.

4.2. INTEGRATION OF QTLs FOR TUBER COLOUR VARIATIONS WITH GENOMIC INFORMATION IN SWEET POTATO

4.2.1. Mining of QTLs

By querying various publications on tuber flesh colour in sweet potato, it was found that only three studies have been done so far on QTL for tuber flesh colour variation in sweet potato. Out of this three works, the data regarding the QTL information on root flesh colour from Chang *et al.* (2009) and Cerventas Flores *et al.* (2011) studies were not available from any of the sources. Information on QTLs identified for β-carotene content from a work conducted by Nair *et al.* (2017) were retrieved for the identification of candidate genes controlling the trait of tuber flesh colour variation in sweet potato.

4.2.2. Identification of tuber flesh colour controlling candidate genes

The flanking markers were taken when available, as queries to perform a BlastN search against the sweet potato genome sequence databases- Ipomoea genome hub and sweet potato genomics resource. The e-PCR search for the marker sequence information of non EST-SSR markers resulted in the primer alignment of only two SSR markers, IB-1809 and IB-242 against the reference genome. Among the thirteen QTLs identified for β-carotene trait in sweet potato both by simple and composite interval mapping, only QTL1, QTL4, QTL2 have flanking sequence markers available. It has been identified that the marker sequences flanking QTL1 on chromosome 5 (chromosome position: 5667728 - 7995922) (Tables. 22&23). Upon the functional annotation of the identified chromosomal region, 70% of the loaded sequences were found to be annotated with top hit species similarity to *Ipomoea nil* (Figures 29&29). QTL4 flanking marker sequence were found on chromosome 10 (chromosome position: 12292746 - 17943538) (Tables 24&25) and 75% of the chromosomal region were annotated with top hit species similarity to *Ipomoea nil* upon functional annotation (Figures 30&31). The flanking marker sequences of QTL2 were found on chromosome 10 (chromosome position: 17659697- 20039937) from a similarity search with *Ipomoea trifida,* are relative species of *Ipomoea batatas.* The genome assembly of *Ipomoea trifida* provided in

sweet potato genomics resource database was used to identify the flanking marker sequence (Table 26). About 69% of the identified chromosomal region were found to be annotated upon functional annotation with huge similarity to *Ipomoea nil* species (Figures 32&33). The genomic sequences are predicted to encode five genes from three QTLs involved mainly in β-carotene biosynthesis. Besides the genes identified for β-carotene biosynthesis, four genes involved in anthocyanin biosynthesis were also identified from the QTL chromosomal region. Dammarenediol II synthase, cytochrome p450 genes involved in carotenoid biosynthesis pathway and caffeoylshikimate esterase, R2R3-MYB transcriptional regulator of anthocyanin biosynthesis were found on QTL1 chromosomal region (Table 27). 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase, hydroxymethylglutaryl-CoA lyase involved in carotenoid biosynthesis and chorismate mutase, 4-coumarate CoA ligase of anthocyanin biosynthesis pathway were identified on QTL4 chromosomal region (Table 28). GGPS, precursor of carotenoid pathway was identified on the chromosomal region of QTL2 (Table 29).

Table 22. Chromosome position of IB 1809 marker sequence on sweet potato genome assembly from ipomoea genome hub database

Query	Identity	Chromosome	Start	End
IB 1809	100.0%	7	28505477	28505500
IB 1809	100.0%	7	25327510	25327533
IB 1809	100.0%	8	37473640	37473661
IB 1809	100.0%	9	2605230	2605251
IB 1809	100.0%	12	46128731	46128775
IB 1809	100.0%	12	28265852	28265876
IB 1809	100.0%	12	28064850	28064872
IB 1809	100.0%	14	28754950	28754979
IB 1809	100.0%	14	12605503	12605529
IB 1809	**100.0%**	**5**	**7995900**	**7995922**

Table 23. Chromosome positionof IB-242marker sequence on sweet potato genome assembly from ipomoea genome hub database

Query	Identity	Chromosome	Start	End
IB-242	100.0%	2	35233528	35233550
IB-242	**100.0%**	**5**	**5667728**	**5667756**
IB-242	100.0%	scaffold26954	1108	1133

Figure 26. Functional annotation result of chromosome region of QTL1

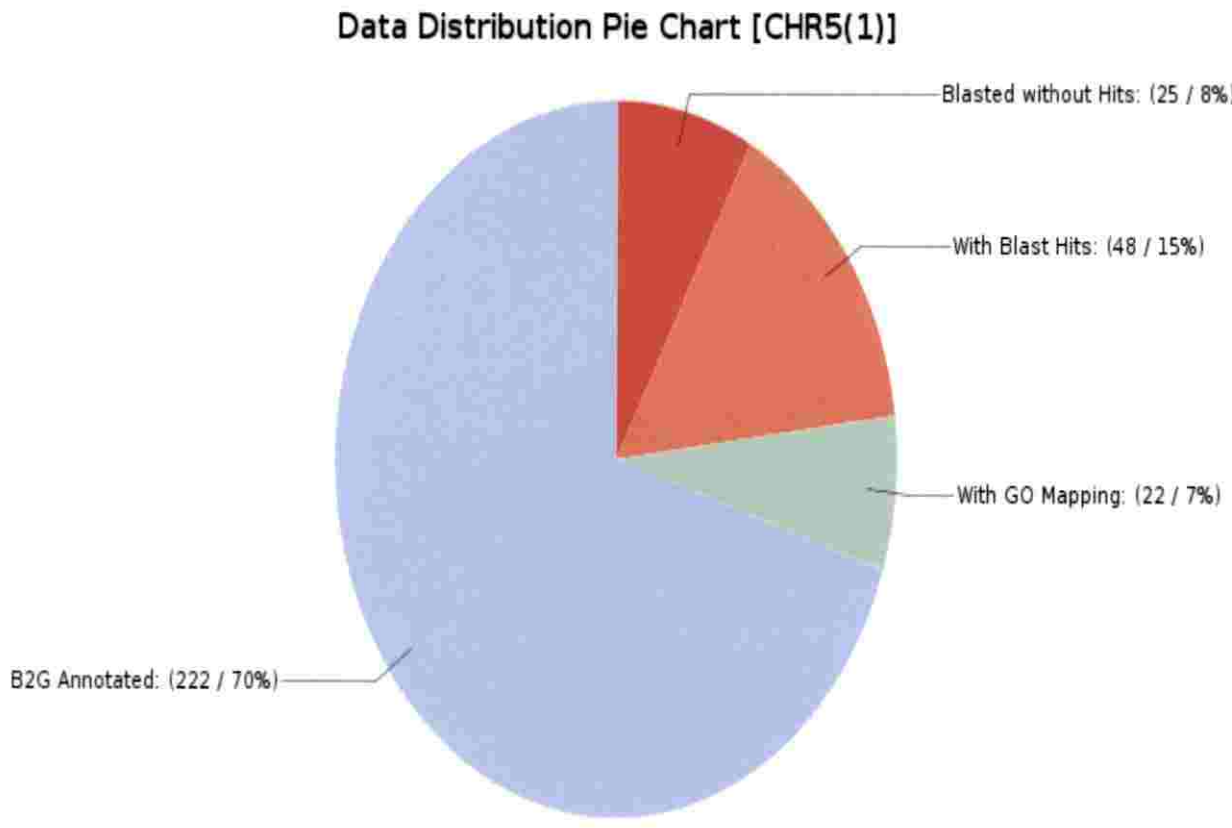


Figure 27. Blast result showing the top hit species distribution of QTL1 chromosomal region

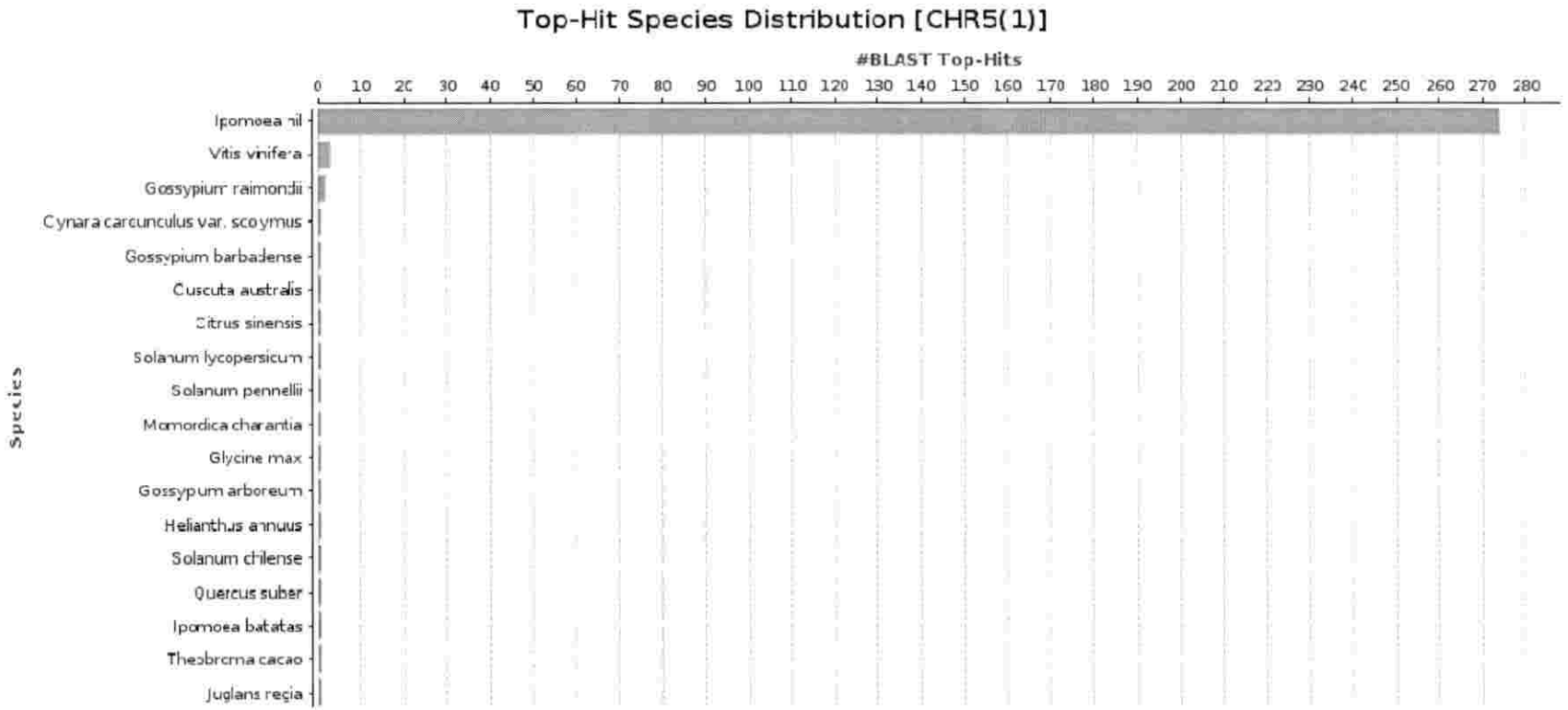


Table24. Chromosome position of markerGDS0134 sequence on sweet potato genome assembly from ipomoea genome hub database

Query	Identity	Chromosome	Start	End
GDS0134	**99.3%**	**10**	**12292746**	**12293018**

Table 25. Chromosome positionof GDS0215marker sequence on sweet potato genome assembly from ipomoea genome hub database

Query	Identity	Chromosome	Start	End
GDS0215	100.0%	scaffold6765	11615	11640
GDS0215	100.0%	10	17943518	17943543
GDS0215	100.0%	scaffold6765	11617	11641
GDS0215	100.0%	7	2856722	2856745
GDS0215	100.0%	10	17940281	17940303
GDS0215	100.0%	7	2856722	2856744
GDS0215	**100.0%**	**10**	**17943519**	**17943538**

Figure 28. Functional annotation result of chromosome region of QTL4

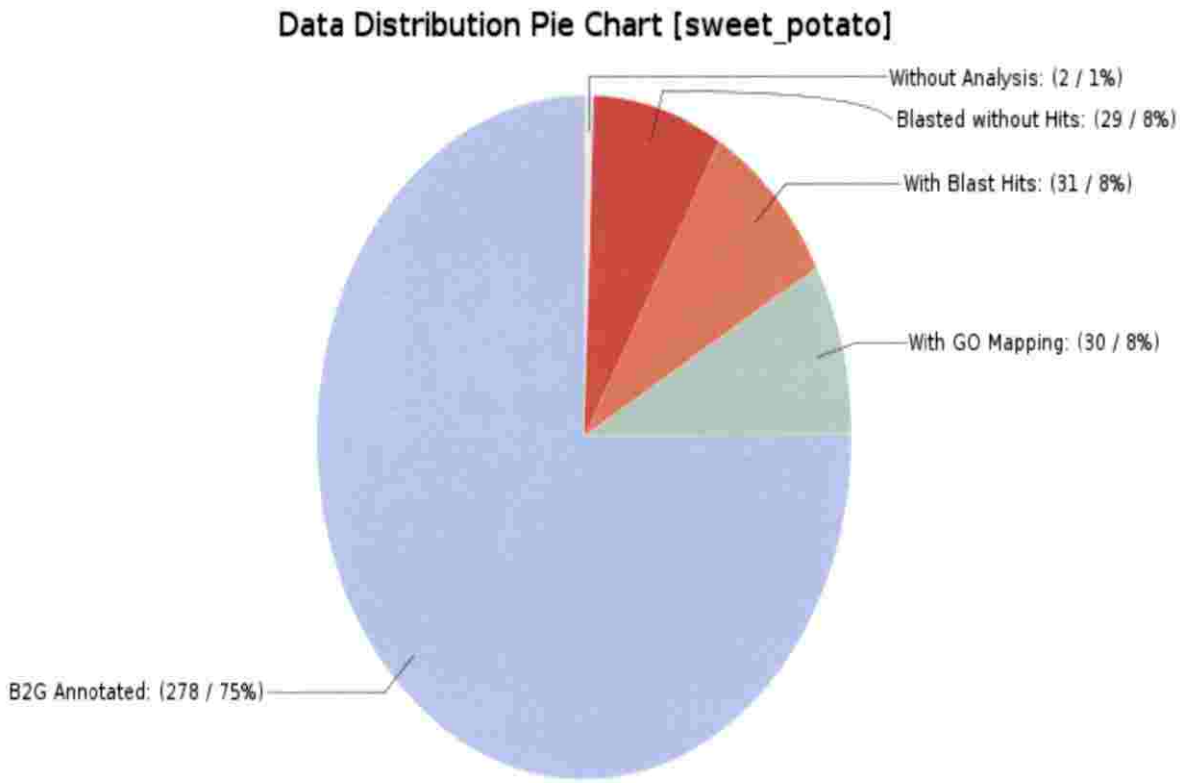


Figure 29. Blast result showing the top hit species distribution of QTL4 chromosomal region

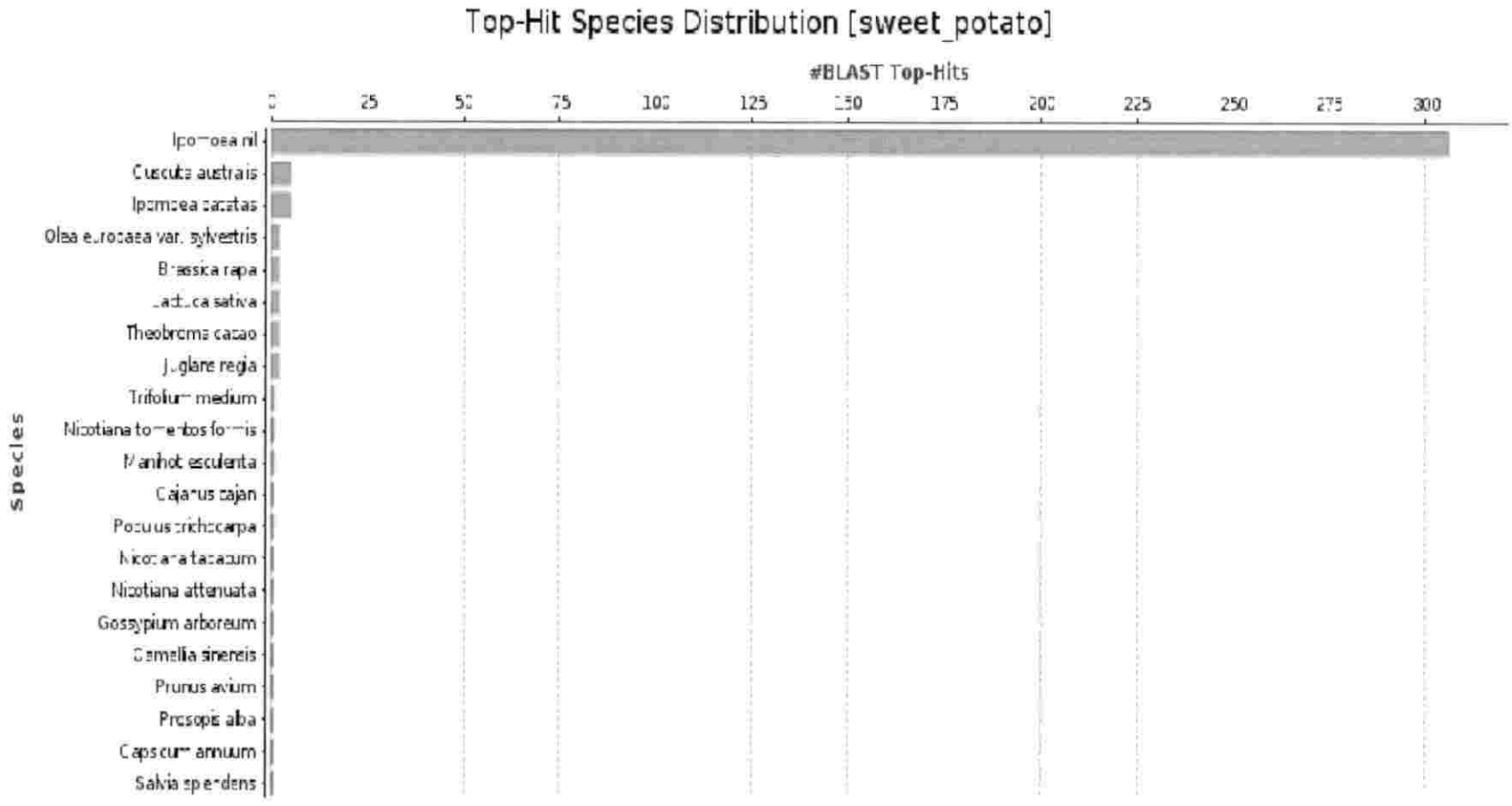


Table 26. Chromosome position of GDS0215 and GDS1059marker sequence on sweet potato genome assembly from sweet potato genomics resource database

Query	Chromosome	Right primer position	Left primer position
GDS0215	**Chr10**	**20039822**	**20039937**
GDS1059	**Chr10**	**17659697**	**17659854**

Figure 30. Functional annotation result of chromosome region of QTL2

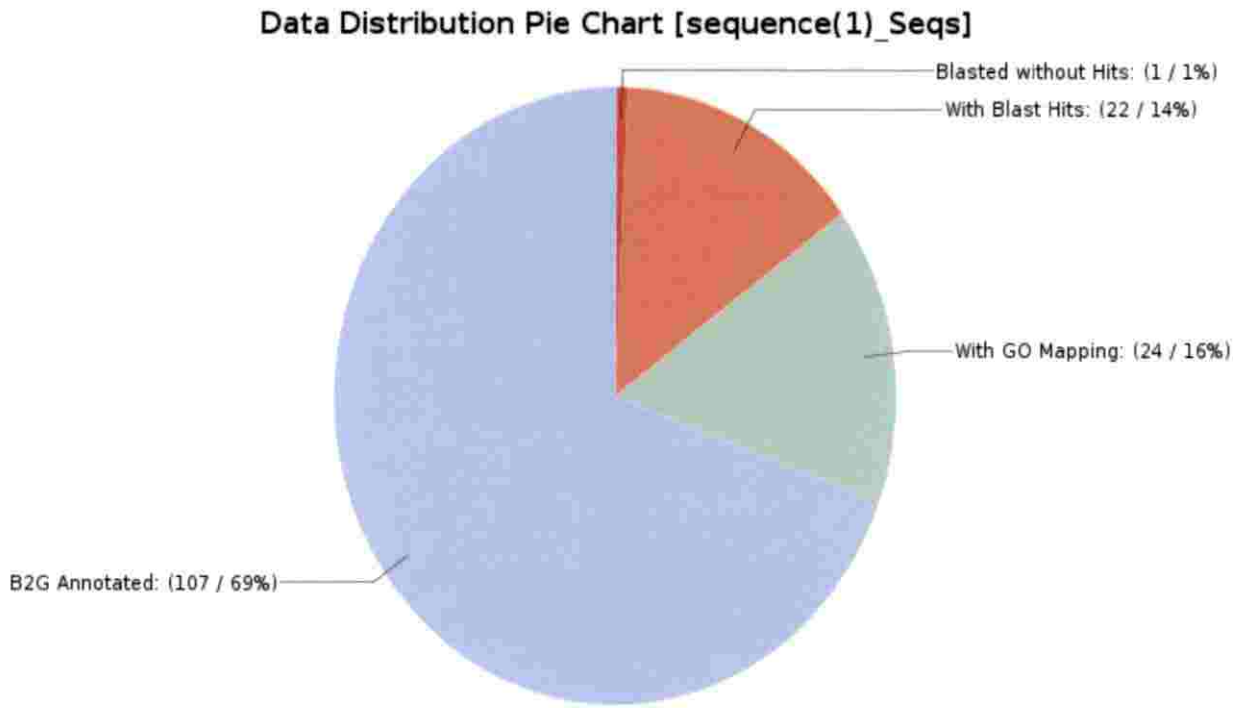


Figure 31. Blast result showing the top hit species distribution of QTL2 chromosomal region

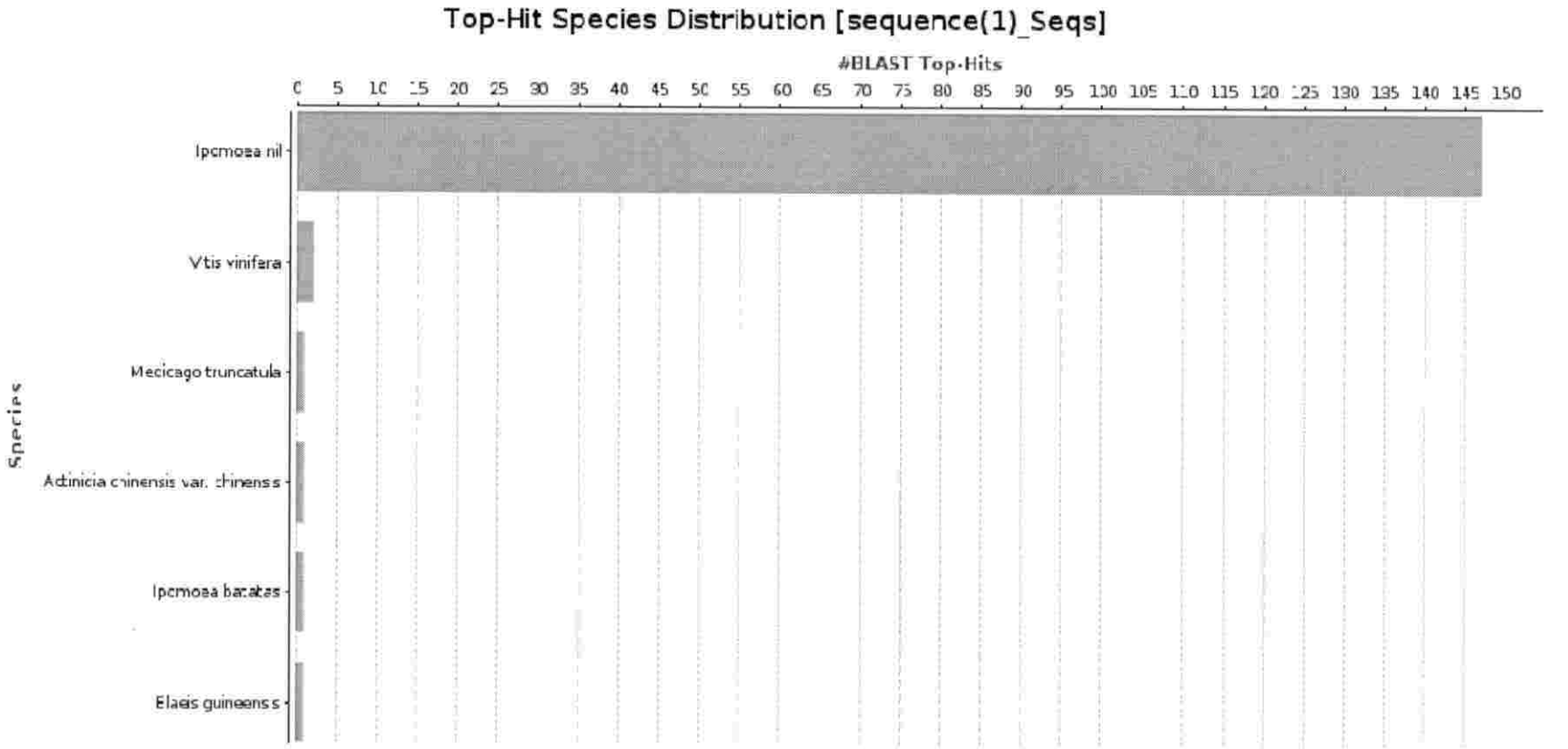


Table 27. List of genes associated with pigment production identified from the QTL1 chromosome region

Query Name	Significant Alignments	Scientific Taxonomy	E-Value	Sim.	Acc.	Gene Name (Taxa Id)	Xref (DB)	G.O. Mapping
G28205|TU46247(0.3)_chr5_7853610-7854138	**PREDICTED: dammarenediol II synthase-like**	*Ipomoea nil*	2.27512e-69	96.2%	XP_019170944	-	-	-
G28212|TU46256(1.0)_chr5_7945842-7946577	**PREDICTED: cytochrome P450 71A1-like**	*Ipomoea nil*	1.41831e-128	96.6%	XP_019151834	-	-	-
G28234|TU46285(0.5)_chr5_7752899-7753771	**PREDICTED: caffeoylshikimate esterase**	*Ipomoea nil*	3.91558e-63	98%	XP_019182898	-	-	-
G28109|TU46077(0.6)_chr5_6531258-6532251	**R2R3-MYB transcriptional regulator**	*Ipomoea nil*	1.67429e-48	96.8%	BAE94710	InMYB3 (35883)	BAE94710 (Interpro)	GO:0003677 GO:0005634

Table 28. List of genes associated with pigment production identified from the QTL4 chromosome region

Query Name	Significant Alignments	Scientific Taxonomy	E-Value	Sim.	Acc.	Gene Name (Taxa Id)	Xref (DB)	G.O. Mapping
G3437|TU5643(16.9)_chr10_15134127-15137365	**PREDICTED: 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase, chloroplastic isoform X2**	*Ipomoea nil*	3.54349e-146	97.3%	XP_019191109	-	-	-
G3560|TU5836(2.2)_chr10_17831903-17836189	**PREDICTED: hydroxymethylglutaryl-CoA lyase, mitochondrial-like isoform X4**	*Ipomoea nil*	0	98.4%	XP_019187439	-	-	-
G3508|TU5756(8.9)_chr10_16425416-16426269	**Chorismate mutase 2-like**	*Papaver somniferum*	2.67854e-48	97%	XP_026414587	-	-	-
G3534|TU5794(5.1)_chr10_17300344-17302516	**4-coumarate:CoA ligase**	*Ipomoea batatas*	2.10952e-151	100%	BAG82851	Ib4CL (4120)	BAG82851 (InterPro)	GO:0003824 GO:0016874

Table 29. List of genes associated with pigment production identified from the QTL2 chromosome region

Query Name	**Significant Alignments**	**Scientific Taxonomy**	**E-Value**	**Sim.**	**Acc.**	**Gene Name (Taxa Id)**	**Xref (DB)**	**G.O. Mapping**
CP025653.1:20010487-20017150	**Geranylgeranyl diphosphate synthase**	*Ipomoea batatas*	0	99.7%	ACF37217	G8XQT1 (4120)	ACF37217 (InterPro)	GO:0008299 GO:0016740

4.2. EXPERIMENTAL VALIDATION

The expression of two randomly selected candidate genes identified from differential expression analysis were detected using the RT product amplified and quantified using 2X H-eff qPCR master mix, Rox qPCR assay.

4.2.1. Candidate genes and primers

Based on the identified candidate genes, GGPS and phytoene synthase specific forward and reverse primers were designed. The primer details are given in Table 30.

4.2.2. Total RNA isolation

Total RNA isolation was performed from the tubers of two sweet potato varieties, Co-34 (white fleshed variety) and Bhu-sona (Orange fleshed variety) available at ICAR-CTCRI using RNeasy plant mini kit of Qiagen in accordance with manufacturer's protocol and lithium chloride method. A distinct or intact RNA with minimum RNA degradation and minimum genomic contamination were observed on agarose gel, showing good quality total RNA extraction. Gel image of RNA isolated from tubers of Co-34 and Bhu-sona is displayed on plate 1.

4.2.3. RNA quantification

The concentration of RNA samples of CO-34 AND Bhu-sona was determined by using a Nano-drop. The concentration and A260/280 of the RNA samples is shown in Table 31.

4.2.4. cDNA synthesis

The isolated RNA samples of white and orange varieties were converted into cDNA using RevertAid First Strand cDNA Synthesis Kit of Thermofisher scientific. The concentraion of cDNA samples were quantified and diluted to a final concentration of 100 ng μl^{-1} and was used for the expression study using real time PCR.

Table 30. Primer details used for validation

Primer	Sequence (5'-3')	Product size(bp)
SP1-F	CACCATGAGGTCGATGAATCTTGT	354
SP1-R	TTCGTTAATTCTGTCTATAAGCTA	354
SP2-F	AAGTTCTTCGACGAGGCTGA	260
SP2-R	GCAGTTTCTTTGGCTTGCTT	260
Actin-F	CCCAAAAGCCAACAGAGAGA	149
Actin-R	CATCACCGAGTCCAACACAAT	149

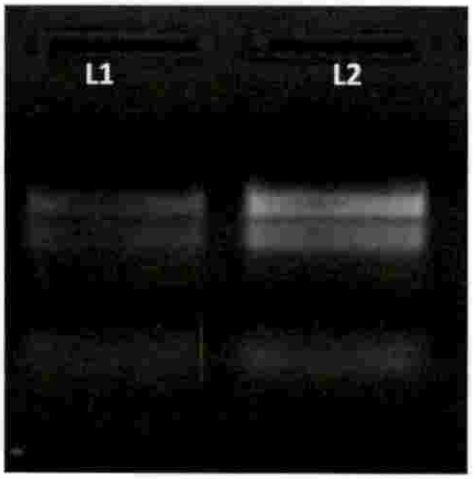


Plate 1. Gel image of RNA isolated from two different tuber varieties of sweet potato Co-34 and Bhu-sona.

L1- Tuber of Co-34, **L2**- Tuber of Bhu-Sona.

Table 31. Concentration and absorbance of isolated RNA

Sample	Concentration (ng/µl)	A260/280
Co-34	537.096	2.54
Bhu-Sona	106.112	2.42

4.2.5. RT-qPCR

For studying gene expression, the genes present in the white and orange fleshed variety were targeted using designed specific primers SP1-F, SP1-R and SP2-F, SP2-R. The SYBR green PCR assay was used for studying gene expression. The relative gene expression level of orange and white fleshed varieties was studied using $2^{-\Delta\Delta CT}$ method. Actin was used as the reference gene for the expression study.

The standard fluorescent amplification representing exponential growth of PCR products was observed in each cycle, yielding threshold cycle (Ct) values. The Ct values is given in the logarithmic scale and inversely proportional to the quantity of cDNA. Thus highly expressed genes have low Ct value and low expressed genes have high Ct value. The fold change (-ΔΔCT) can be calculated by comparing the normalized expression (ΔCt) of the two conditions. The fold change, *viz.* the expression ratio, indicated the upregulation and downregulation of the gene. The orange tuber variety (Bhu-sona) had a fold change of about 3.0 by amplifying with the primer SP-1 specific to the gene geranyl geranyl pyrophosphate and didn't showed any fold change by amplifying with the primer SP-2 specific to phytoene synthase (Figure 32).

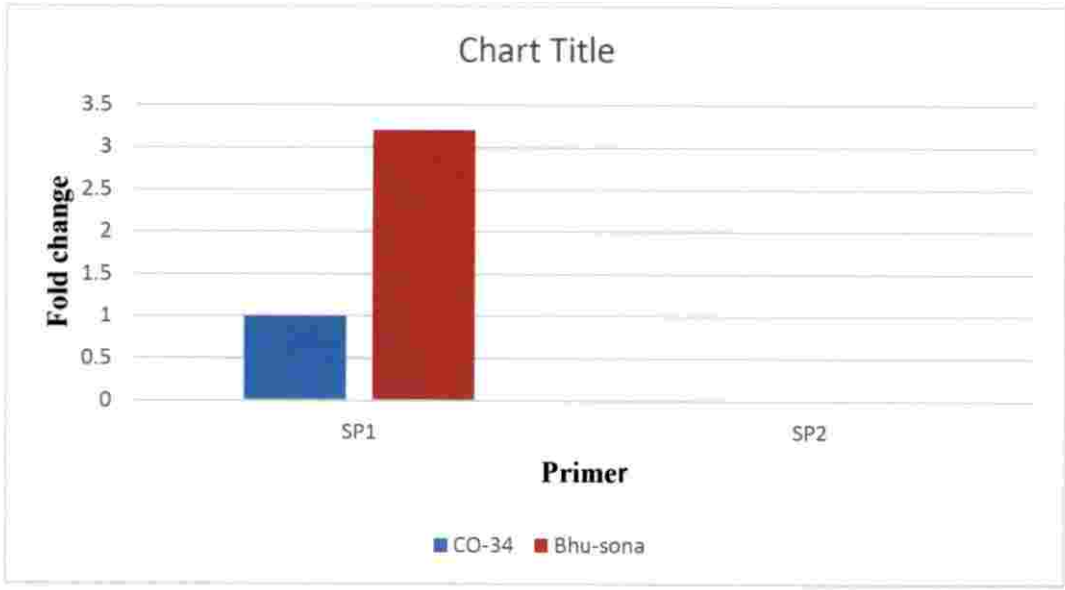


Figure 32. Relative gene expression of orange variety (Bhu-sona) and white variety (Co-34) variety.

DISCUSSION

5. DISCUSSION

The study entitled "Integration of Quantitative Trait Locus (QTL) for tuber colour variations with genomic information in sweet potato (*Ipomoea batatas* L.) was conducted to identify the differentially expressed genes for various tuber colours in sweet potato and to identify the candidate genes for tuber flesh colour variation by the integration of QTL with genomic information of sweet potato.

The quality improvement of sweet potato has been focused on breeding different types of cultivars with high content of starch, carotenoids and anthocyanins. Improvement in the content of the functional components is an important target for breeding. As sweet potato is an autohexaploid with self and cross compatibility, its genetic analysis has been difficult. The identification of candidate genes involved in the accumulation of functional components, carotenoids and anthocyanins will facilitate the use of these candidate genes to improve the efficiency of cultivar development.

Carotenoids are widely distributed pigments in plants and play an important role as light harvesting pigments in photosynthetic organisms. Carotenoids are derived from isopentenyl diphosphate (IPP) which is synthesized mainly by two independent pathways in higher plants-methyl-erythritol phosphate (MEP) pathway and mevalonate (MVA) pathway (Flores-perez *et al.*, 2010). Four molecules of IPP are converted to geranyl geranyl diphosphate (GGPP) by IPP isomerase. The first step in the carotenoid pathway, involves the catalyzation of phytoene synthase by the condensation of two molecules of GGPP to produce phytoene and subsequent production of all trans-lycopene from phytoene. Lycopene is the substrate for two competing cyclases, lycopene epsilon cyclase and lycopene β cyclase. Lycopene is processed by two carotenoids biosynthetic pathway, the α- branch pathway and the β branch pathway. The production of trans-lycopene from phytoene requires a complex set of four reactions involving phytoene desaturase (PDS), zeta-carotene isomerase, zeta-carotene desaturase and carotenoid isomerase. In β-carotene branch pathway, hydroxylation of β-hydroxylase converts β-carotene to zeaxanthin and then zeaxanthin epoxidase

(ZEP) mediates the formation of violaxanthin. Violaxanthin is converted to neoxanthin by neoxanthin synyhase. In higher plants, the carotenoid metabolic pathway and the function of the biosynthetic enzymes have been well studied and series of genes involved in the pathway have been cloned and characterized.

The identification of candidate genes involved in complex biosynthetic pathways is very difficult due to the limited availability of genomic data in non-model plants. High throughput transcriptome sequencing can generate large amount of data on genome wide transcriptome.In the present study, transcriptomes of six tuber varieties of sweet potato with varying flesh colours of white (Tuber1 and Tuber2), purple (Tuber3 and Tuber4) and orange (Tuber5 and Tuber6) sequenced using Nextseq 500 platform were used as the preliminary data. 26,081,266 and 23,436,464 high quality reads were produced from white fleshed variety ofTuber1 and Tuber2 respectively. 24,506,748 and 23,121,241 high quality reads were produced from purple fleshed variety of Tuber3 and Tuber4 respectively whereas 23,835,350 and 25,841,424 high quality reads were obtained from Tuber5 and Tuber6 of orange fleshed variety. The high quality reads obtained for orange fleshed tuber varieties (Tuber5 and Tuber6) were found to be longer than that reported from Weiduoli, an orange fleshed variety and its mutant HVB-3 with high β- carotene content. 13,767,387and 9,837.090 high quality reads were obtained from Weiduoli and HVB-3 respectively (Li *et al.*, 2015). A total of 37,4813 unigenes with were harvested from the six tuber varieties. A total of 35,909 unigenes with an average length of 533bp and an N50 of 669bp were obtained from Weiduoli, an orange fleshed sweet potato and its mutant HVB-3 with high β-carotene content (Li *et al.*, 2015). Thus the unigenes obtained in our study for the orange fleshed tuber varieties (Tuber5 and Tuber6) were found to be larger than that reported in Weiduoli and HVB-3.

White, orange and purple fleshed varieties of sweet potato were used for studying the genes that participated in carotenoid biosynthesis pathway. Based on the RNA seq data of white (Tuber1 and Tuber2) and orange (Tuber5 and Tuber6) 22,534 out of the 111,231 transcripts were variably expressed between white and orange

varieties. Among them, 5472 were upregulated and 17,062 were downregulated in orange compared to white. From orange and purple (Tuber3 and Tuber4) RNA seq data, 27,431 out of the 111,231 transcripts were found to be differentially expressed between orange and purple varieties with 11,670 upregulated genes and 15,761 downregulated genes in orange compared to purple. RNA seq data of purple and white produced 22,590 variably expressed genes out of 111,231 transcripts between purple and white varieties. Among them, 7,622 were upregulated and 14,968 were downregulated in purple compared to white. Li *et al.* (2015) reported 874 differentially expressed genes between Weiduoli, an orange fleshed variety and its mutant HVB-3 with high β- carotene content, 401 of which were upregulated and 473 were downregulated in HVB-3 compared to Weiduoli. In another study, the RNA-Seq database of MN1 (mutant of Ningzishu 1, purple flesh sweet potato) and WTN1 (wild type of Ningzishu 1), 7, 627 out of the 88,509 unigenes, were variably expressed between MN1 and WTN1. Among them, 2,995 were up-regulated, and 4,632 were down-regulated (Ma *et al.*, 2016). In the present study instead of identifying the differentially expressed genes from unigenes, transcripts were analysed for their variable expression between different varieties. As the transcripts sequences were analysed, probability of losing information on multigene families or isoforms with high-sequence similarities will be less compared to that of using unigenes obtained after elimination of redundant sequences.

The present study after enrichment analysis of the differentially expressed genes obtained for each of the pairwise analysis, showed terms isoprenoid biosynthetic and metabolic process enriched among the differentially expressed genes for orange and white pairwise analysis,terpenoid metabolic process enriched among the differentially expressed genes for orange and purple pairwise analysis and antioxidant activity enriched among the differentially expressed genes for purple and white pairwise analysis. The functional annotation of the twenty-four transcripts associated with the gene ontology term isoprenoid biosynthetic process showed six genes involved in the carotenoid biosynthesis pathway. Upon the functional annotation of twenty-nine transcripts for the term isoprenoid metabolic

process resulted in identification of seven genes in carotenoid biosynthesis pathway. Thirteen transcripts were associated with terpenoid metabolic process resulted in the identification of three genes mainly involved in carotenoid biosynthesis pathway. Antioxidant activity related fourteen transcripts upon functional annotation showed genes with peroxidase activity. The present results showed seven differentially expressed genes related to carotenoid biosynthesis existed between orange and white, three differentially expressed genes related to carotenoid biosynthesis existed between orange and purple.

The variable genes identified for carotenoid biosynthesis includes β-carotene hydroxylase, zeta-carotene desaturase, Isopentenyl diphosphate isomerase (IPI), squalene synthase, geranyl geranyl diphosphate synthase, zeaxanthin epoxidase and phytoene synthase. β- carotene hydroxylase is the key regulatory enzyme in β branch of carotenoid biosynthesis. Down regulation of the β-carotene hydroxylase gene increases β-carotene and total carotenoids in transgenic cultured cells of sweet potato (Kim *et al.*, 2012). Isopentenyl diphosphate isomerase catalyzes the interconversion of Isopentenyl pyrophosphate (IPP) and Dimethylallyl pyrophosphate (DMAPP). In MEP pathway, both IPP and DMAPP are produced in the terminal branching step by the action of Isopentenyl diphosphate synthase (Rohdich, 2003). This conclude that IPI is required for the full function of the pathway. zeta-carotene desaturase is involved in the steps which sequentially convert 9,9'-di-*cis*-zeta-carotene to pro-lycopene and to all-*trans*-lycopene. The expression level of zeta carotene was reported as increasing during the fruit development and ripening of watermelon (Grassi *et al.*, 2013). Phytoene synthase (PSY) is a key regulator in the carotenoid biosynthetic pathway. The expression of phytoene synthase was reported as increasing in the flesh and peel of papaya demonstrated that phytoene synthase was active in the carotenoid biosynthetic pathway in papaya (Shen *et al.*, 2019). The formation of geranyl geranyl pyrophosphate (GGPP) is a key step in biosynthetic pathway of terpenes and carotenoids. A cDNA namely IbGGPS was cloned from storage roots of sweet potato and its over expression in Arabidopsis increased the contents of total carotenoids (Chen *et al.*, 2015). In a study, the role of zeaxanthin epoxidase in the

synthesis and accumulation of carotenoids in *B.rapa* has been reported (Tuan *et al*., 2012).

Computational approaches have been utilized for the prediction of candidate genes of targeted QTLs based on the availability of genomic sequences and markers linked to the QTL. In the present study five genes involved in the carotenoid biosynthesis pathway and four genes in the anthocyanin biosynthesis pathway were identified from three QTLs controlling the β-carotene trait. Dammarenediol II synthase, cytochrome P450, 2-C-Methyl-D-erythritol 4-phosphate cytidylyltransferase, hydroxymethylglutaryl-CoA lyase and Geranylgeranyl diphosphate synthase were the candidate genes identifies for the carotene biosynthesis pathway. *P450CYP707A* encoding ABA 8′-hydroxylases and *LUT1* encoding cytochrome P450-type monooxygenase (CYP97C1) have been proved to regulate carotenoid biosynthesis in *Arabidopsis* (Kushiro *et al*., 2004; Tian *et al*., 2004). The biosynthesis of a terpenoid backbone resulted in a universal building block, isopentenyl pyrophosphate (IPP) is generated from the 2-*C*-Methyl-D-erythritol 4-phosphate (MEP) pathway in plastids (Hemmerlin *et al*., 2012). IPP was biosynthesized from Acetyl-CoA by sequential actions of many enzymes including hydroxymethylglutaryl-CoA lyase. Dammarenediol II synthase is 2,3-oxidosqualene in which the biosynthesis of squalene from two farnesyl diphosphates (FPPs), catalyzed by squalene synthase is the first committed step in carotenoid biosynthesis pathways.

Besides the candidate gene involved in the carotenoid pathway, genes in the anthocyanin biosynthesis pathway has also been identified. This include caffeoylshikimate esterase, R2R3-MYB transcriptional regulator, chorismate mutase 2-like and 4-coumarate CoA ligase. MYB transcriptional factors have been identified in various plant species as regulators of flavonoid biosynthesis in flowers, seeds, and fruits. Mano *et al*. (2007) isolated the MYB genes *IbMYB1* and *IbMYB2* from a purple-fleshed sweet potato and suggested that *IbMYB1* controls anthocyanin biosynthesis specifically in the flesh of sweet potato storage roots. Chorismate mutase is involved in the catalyzation of

chemical reaction for the converwsion of chorismate to the prephenate in the shikimate pathway. The anthocyanin biosynthetic pathway starts with the chalcone synthase (CHS) mediated synthesis of naringenin chalcone from 4-coumaroyl-CoA and malonyl-CoA. Caffeoylshikimate esterase catalyze the hydroxylase of the *meta*-position of the hydroxycinnamoyl shikimate involved in the anthocyanin biosynthesis pathway. The two candidate genes GGPS and phytoene synthase identified as differentially expressed between orange and white fleshed varieties were validated using q-PCR. The discovery and characterization of genes involved in the carotenoid pathway through transcriptome analysis can be utilized for improvement of sweet potato by the introduction of the genes to commercial sweet potato cultivars. Fine mapping of QTL controlling β-carotene trait will facilitate Marker Assisted Selection.

SUMMARY

6. SUMMARY

The study entitled Integration of Quantitative Trait Loci (QTL) for tuber colour variations with genomic information in sweet potato (*Ipomoea batatas* L.) was conducted at section of extension and social sciences, ICAR-CTCRI. The main objective of the study was to identify the differentially expressed genes for various tuber colours in sweet potato using RNA sequenced data; to integrate QTL information on tuber colour with genomic information in sweet potato and to validate the identified candidate genes.

The raw RNA sequenced data of the six tuber varieties and their replicates were analysed for their quality using the FastQC and high quality clean reads were obtained after processing the RNA sequenced data using trimmomatic. The filtered high quality reads of the samples were *de novo* assembled into transcripts using Trinity RNA seq assembler. OmicsBox was used for the transcript level quantification for estimating the gene and isoform expression levels from RNA seq data, pairwise expression analysis, enrichment analysis and functional annotation of the enriched sequences. Based on the RNA seq data of white (Tuber1 and Tuber2) and orange (Tuber5 and Tuber6) 22,534 out of the 111,231 transcripts were variably expressed between white and orange varieties. Among them, 5472 were upregulated and 17,062 were downregulated in orange compared to white. From orange and purple (Tuber3 and Tuber4) RNA seq data, 27,431 out of the 111,231 transcripts were found to be differentially expressed between orange and purple varieties with 11,670 upregulated genes and 15,761 downregulated genes in orange compared to purple. RNA seq data of purple and white produced 22,590 variably expressed genes out of 111,231 transcripts between purple and white varieties. Among them, 7,622 were upregulated and 14,968 were downregulated in purple compared to white. The present results showed seven differentially expressed genes related to carotenoid biosynthesis existed between orange and white, three differentially expressed genes related to carotenoid biosynthesis existed between orange and purple. The variable genes identified for carotenoid biosynthesis includes β-carotene hydroxylase, zeta-

carotene desaturase, Isopentenyl diphosphate isomerase, squalene synthase, geranyl geranyl diphosphate synthase, zeaxanthin epoxidase and phytoene synthase.

The QTLs involved in tuber flesh colour in sweet potato were retrieved from QTL studies on tuber colour variation. To link genetic map with genomic resource, sequence alignment of marker sequences was performed with sweet potato genome assemblies available at Ipomoea Genome hub and sweet potato genomics resource database. The chromosome location of QTL was identified by performing similarity searches (blastn) and further functional annotation of the identified chromosomal region was done using OmicsBox. The functional annotation of the chromosomal region was done to predict the candidate gene for tuber flesh in sweet potato. In the present study five genes involved in the carotenoid biosynthesis pathway were identified from three QTLs controlling the β-carotene trait. Dammarenediol II synthase, cytochrome P450, 2-C-Methyl-D-erythritol 4-phosphate cytidylyltransferase, hydroxymethylglutaryl-CoA lyase and Geranylgeranyl diphosphate synthase were the candidate genes identified for the carotene biosynthesis pathway. The candidate genes identified was validated by using an accession of white and orange fleshed sweet potato variety. The genes associated with carotenoid biosynthesis in storage roots may enable efficient breeding of varieties with high provitamin A content. These resources will facilitate genome enabled breeding in the important food security crop.

REFERENCES

7. REFERENCE

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. and Harris, M.A. 2000. Gene ontology: tool for the unification of biology. *Nat. genet.* 25(1): 25.

Austin, D.F. 1988. The taxonomy, evolution and genetic diversity of sweet potatoes and related wild species. *Exploration, maintenance, and utilization of sweet potato genetic resources*: 27-60.

Ben-Amotz, A. and Fishier, R. 1998. Analysis of carotenoids with emphasis on 9-cis β-carotene in vegetables and fruits commonly consumed in Israel. *Food Chem.* 62(4): 515-520.

Bender, D. 2002. *An introduction to nutrition and metabolism*. CRC Press.

Botella–Pavía, P. and Rodríguez- Concepción, M. 2006. Carotenoid biotechnology in plants for nutritionally improved foods. *Physiologia Plant.*126(3): 369-381.

Bovell–Benjamin, A.C. 2007. Sweet potato: a review of its past, present, and future role in human nutrition. *Adv. Food. Res*. 52: 1-59.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. mol biol.* 268(1): 78-94.

Buteler, M.I., Jarret, R.L. and LaBonte, D.R. 1999. Sequence characterization of microsatellites in diploid and polyploid *Ipomoea. Theor. and Appl. Genet.* 99(1-2): 123-132.

Cervantes-Flores, J.C., Sosinski, B., Pecota, K.V., Mwanga, R.O.M., Catignani, G.L., Truong, V.D., Watkins, R.H., Ulmer, M.R. and Yencho, G.C. 2011. Identification of quantitative trait loci for dry-matter, starch, and β-carotene content in sweet potato. *Mol. Breed.* 28(2): 201-216.

Cervantes-Flores, J.C., Yencho, G.C., Kriegner, A., Pecota, K.V., Faulk, M.A., Mwanga, R.O. and Sosinski, B.R. 2008. Development of a genetic linkage map and identification of homologous linkage groups in sweet potato using multiple-dose AFLP markers. *Mol. Breed.* 21(4): 511-532.

Chang, K.Y., Lo, H.F., Lai, Y.C., Yao, P.J., Lin, K.H. and Hwang, S.Y. 2009. Identification of quantitative trait loci associated with yield-related traits in sweet potato (*Ipomoea batatas*). *Bot. stud.* 50(1): 43-55.

Chen, B.H. and Chen, Y.Y. 1993. Stability of chlorophylls and carotenoids in sweet potato leaves during microwave cooking. *J. of agric. and food chem.* 41(8): 1315-1320.

Chen, W., He, S.Z., Liu, D.G., Patil, G.B., Zhai, H., Wang, F.B., Stephenson, T.J.,Wang, Y.N., Wang, B., and Valliyodan B. 2015.A sweet potato geranylgeranyl pyrophosphate synthase gene, *IbGGPS*, increases carotenoid content and enhances osmotic stress tolerance in *Arabidopsis thaliana*. *PLoS ONE10*: e0137623

Cunningham Jr, F.X. and Gantt, E. 1998. Genes and enzymes of carotenoid biosynthesis in plants. *Ann.rev. of plant biol* 49(1): 557-583.

deMiguel, M., Cabezas, J.A., de María, N., Sánchez-Gómez, D., Guevara, M.Á., Vélez, M.D., Sáez-Laguna, E., Díaz, L.M., Mancha, J.A., Barbero, M.C. and Collada, C. 2014. Genetic control of functional traits related to photosynthesis and water use efficiency in Pinus pinaster Ait. drought response: integration of genome annotation, allele association and QTL detection for candidate gene identification. *BMC genomics* 15(1): 464.

De Pee, S., Bloem, M.W., Gorstein, J., Sari, M., Yip, R. and Shrimpton, R. 1998. Reappraisal of the role of vegetables in the vitamin A status of mothers in Central Java, Indonesia. *Am. J. clin. Nutr.* 68(5): 1068-1074.

Deshmukh, R., Singh, A., Jain, N., Anand, S., Gacche, R., Singh, A., Gaikwad, K., Sharma, T., Mohapatra, T. and Singh, N. 2010. Identification of candidate

genes for grain number in rice (Oryza sativa L.). *Funct. integrated genomics* 10(3): 339-347.

Diop, A. and Calverley, D.J.B. 1998. Storage and Processing of Roots and Tubers in the Tropics. *Food and Agriculture Organization of the United Nations, Agro-industries and Post-Harvest Management Service, Agricultural Support Systems Division.*

FAO [Food and Agriculture Organisation]. 2014.

FAO [Food and Agriculture Organisation]. 2015

FAO [Food and Agriculture Organisation]. 2016.

FAOSTAT [Food and Agriculture Organaization Corporate Statistical Database] 2016.

FAOSTAT [Food and Agriculture Organaization Corporate Statistical Database] 2019

Firon, N., LaBonte, D., Villordon, A., Kfir, Y., Solis, J., Lapis, E., Perlman, T.S., Doron-Faigenboim, A., Hetzroni, A., Althan, L. and Nadir, L.A. 2013. Transcriptional profiling of sweet potato (Ipomoea batatas) roots indicates down-regulation of lignin biosynthesis and up-regulation of starch biosynthesis at an early stage of storage root formation. *BMC genomics* 14(1): 460.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S. and Gil, L. 2013. Ensembl 2014. *Nucl. acids res.* 42:.749-755.

Flores-Perez, U., Perez-Gil, J., Closa, M., Wright, L.P., Botella-Pavia, P., Phillips, M.A., Ferrer, A., Gershenzon, J. and Rodriguez-Concepcion, M. 2010. Pleiotropic regulatory locus 1 (PRL1) integrates the regulation of sugar responses with isoprenoid metabolism in *Arabidopsis*. *Mol. Plant.* 3(1): 101-112.

Fraser, P.D. and Bramley, P.M. 2004. The biosynthesis and nutritional uses of carotenoids. *Prog. in lipid res.* 43(3): .228-265.

Freyre, R., Iwanaga, M. and Orjeda, G. 1991. Use of *Ipomoea trifida* (HBK.) G. Don germ plasm for sweet-potato improvement. 2. Fertility of synthetic hexaploids and triploids with 2 n gametes of I. trifida, and their interspecific crossability with sweet potato. *Genome*, 34(2): 209-214.

Gasura, E., Mashingaidze, A.B. and Mukasa, S.B., 2008. Genetic variability for tuber yield, quality, and virus disease complex traits in Uganda sweet potato germplasm. *African Crop Sci. J.* 2:*16*.

Gelli, M., Konda, A.R., Liu, K., Zhang, C., Clemente, T.E., Holding, D.R. and Dweikat, I.M., 2017. Validation of QTL mapping and transcriptome profiling for identification of candidate genes associated with nitrogen stress tolerance in sorghum. *BMC plant biol.* 17(1): 123.

Girard, A.W., Grant, F., Watkinson, M., Okuku, H.S., Wanjala, R., Cole, D., Levin, C. and Low, J. 2017. Promotion of orange-fleshed sweet potato increased vitamin A intakes and reduced the odds of low retinol-binding protein among postpartum Kenyan women. *J. Nutr.* 147(5): 955-963.

Grassi, S., Piro, G., Lee, J.M., Zheng, Y., Fei, Z., Dalessandro, G., Giovannoni, J.J. and Lenucci, M.S., 2013. Comparative genomics reveals candidate carotenoid pathway regulators of ripening watermelon fruit. *BMC genomics,* 14(1): 781.

Gross, S.S. and Brent, M.R. 2006. Using multiple alignments to improve gene prediction. *J. comp. biol.* 13(2): 379-393.

Guimarães, E.P. 2007. Marker-assisted selection: current status and future perspectives in crops, livestock, forestry and fish. *Food & Agriculture Org.*

Gurmu, F., Hussein, S. and Laing, M. 2014. The potential of orange-fleshed sweet potato to prevent vitamin A deficiency in Africa. *Int. J. Vitam. Nutr. Res.* 84(1-2): 65-78.

Harushima, Y., Yano, M., Shomura, A., Sato, M., Shimano, T., Kuboki, Y., Yamamoto, T., Lin, S.Y., Antonio, B.A., Parco, A. and Kajiya, H. 1998. A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genet.* 148(1): 479-494.

Hemmerlin, A., Harwood, J.L., and Bach, T.J. 2012 A raison d'etre for two distinct pathways in the early steps of plant isoprenoid biosynthesis. *Prog. Lipid Res.* 51: 95–148.

Huang, D.W., Sherman, B.T. and Lempicki, R.A., 2008. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucl. acids res.* 37(1): 1-13.

Ishiguro, K., Toyama, J., Islam, M.S., Yoshimoto, M., Kumagai, T., Kai, Y., Nakazawa, Y. and Yamakawa, O. 2004. Suioh, a new sweet potato cultivar for utilization in vegetable greens. *Acta Hortic*: 339-346.

Ishiguro, K., Yoshinaga, M., Kai, Y., Maoka, T. and Yoshimoto, M. 2010. Composition, content and antioxidative activity of the carotenoids in yellow-fleshed sweet potato (*Ipomoea batatas* L.). *Breed.Sci.* 60(4): 324-329.

Islam, S. 2006. Sweet potato (*Ipomoea batatas* L.) leaf: its potential effect on human health and nutrition. *J. Food Sci.* 71(2): 13-121.

Jalal, F.N.M.C., Nesheim, M.C., Agus, Z., Sanjur, D. and Habicht, J.P., 1998. Serum retinol concentrations in children are affected by food sources of beta-carotene, fat intake, and anthelmintic drug treatment. *Am. J. clin. nutr.*68(3): 623-629.

Jian, H., Zhang, A., Ma, J., Wang, T., Yang, B., Shuang, L.S., Liu, M., Li, J., Xu, X., Paterson, A.H. and Liu, L. 2019. Joint QTL mapping and transcriptome sequencing analysis reveal candidate flowering time genes in Brassica napus L. *BMC genomics,* 20(1): 21.

Just, B.J., Santos, C.A.F., Fonseca, M.E.N., Boiteux, L.S., Oloizia, B.B. and Simon, P.W.2007. Carotenoid biosynthesis structural genes in carrot (*Daucus*

carota): isolation, sequence-characterization, single nucleotide polymorphism (SNP) markers and genome mapping. *Theor. Appl. Genet*. 114(4): 693-704.

Kai, Y. 2004. Ayakomachi: New sweet potato cultivar for cooking material and table use. *Sweet potato Res. Front*. 17: 4.

Kai, Y., Katayama, K., Sakai, T. and Yoshinaga, M. 2010. Beniharuka: a new sweet potato cultivar for table use. *Sweet potato. Res. Front.* (23): 420-488

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 sweet potato. *Breed Sci.* :16118.

Kanehisa, M. and Goto, S., 2012. KEGG: Kyoto encyclopedia of genes and genomes. Kanehisa laboratories.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M. and Harte, R.A. 2013. The UCSC genome browser database: 2014 update. *Nucl. acids res.* 42: 764-770.

Kim, S.H., Ahn, Y.O., Ahn, M.J., Jeong, J.C., Lee, H.S. and Kwak, S.S. 2013. Cloning and characterization of an Orange gene that increases carotenoid accumulation and salt stress tolerance in transgenic sweet potato cultures. *Plant physiol. and biochem.* 70: 445-454.

Kim, S.H., Ahn, Y.O., Ahn, M.J., Lee, H.S. and Kwak, S.S. 2012. Down-regulation of β-carotene hydroxylase increases β-carotene and total carotenoids enhancing salt stress tolerance in transgenic cultured cells of sweet potato. *Phytochem.*74: 69-78.

Kim, S.H., Jeong, J.C., Park, S., Bae, J.Y., Ahn, M.J., Lee, H.S. and Kwak, S.S., 2014. Down-regulation of sweet potato lycopene β-cyclase gene enhances tolerance to abiotic stress in transgenic calli. *Mol. Biol.rep.* 41(12): 8137-8148.

Koech, R.K., Malebe, P.M., Nyarukowa, C., Mose, R., Kamunya, S.M., Joubert, F. and Apostolides, Z. 2019. Functional annotation of putative QTL associated with black tea quality and drought tolerance traits. *Sci. Rep.* 9(1): 1465.

Kohlmeier, L. and Hastings, S.B. 1995. Epidemiologic evidence of a role of carotenoids in cardiovascular disease prevention. *Am. J. clin. Nutr.* 62(6): 1370-1376S

Kotera, M., Hirakawa, M., Tokimatsu, T., Goto, S. and Kanehisa, M. 2012. The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. In *Next Generation Microarray Bioinformatics*: 19-39.

Kriegner, A., Cervantes, J.C., Burg, K., Mwanga, R.O. and Zhang, D., 2003. A genetic linkage map of sweet potato [Ipomoea batatas (L.) Lam.] based on AFLP markers. *Mol. Breed.*1(3): 169-185.

Kurabachew, H. 2015. Review on the Impact of Orange Fleshed Sweet Potato on the Reduction of Vitamin A Deficiency under Five Years Old Children and Lactating Women in Ethiopia. *Eur. J. Nutr. & Food Saf.*: 1005-1006

Kushiro, T., Okamoto, M., and Nakabayashi, K. 2004. The *Arabidopsis* cytochrome P450 CYP707A encodes ABA 8′-hydroxylases: key enzymes in ABA catabolism. *The EMBO J.* 23(7):1647–1656.

Li, A., Qingchang, L. and Qingmei, W. 2010. Mapping QTLs for starch content in sweet potato. *Mol. Plant Breed.* 31: 432-487

Li, H., Zhao, N., Yu, X., Liu, Y., Zhai, H., He, S., Li, Q., Ma, D. and Liu, Q. 2014. Identification of QTLs for storage root yield in sweet potato. *Sci Hortica.* 170: 182-188.

Li, R., Zhai, H., Kang, C., Liu, D., He, S. and Liu, Q. 2015. De novo transcriptome sequencing of the orange-fleshed sweet potato and analysis of

differentially expressed genes related to carotenoid biosynthesis. *Int. J. Genomics,* 40: 650-700

Liu, F., Yang, Y., Gao, J., Ma, C. and Bi, Y. 2018. A comparative transcriptome analysis of a wild purple potato and its red mutant provides insight into the mechanism of anthocyanin transformation. *PloS one* 13(1): e0191406.

Liu, Y.S., Gur, A., Ronen, G., Causse, M., Damidaux, R., Buret, M., Hirschberg, J. and Zamir, D. 2003. There is more to tomato fruit colour than candidate carotenoid genes. *Plant biotechnol. J.* 1(3): 195-207.

Low, J.W. 1997. Combating Vitamin A Deficiency Through the Use of Sweet Potato: Results from Phase I of an Action Research Project in South Nyanza, Kenya. International Potato Center.

Low, J.W., Mwanga, R.O., Andrade, M., Carey, E. and Ball, A.M. 2017. Tackling vitamin, A deficiency with biofortified sweet potato in sub-Saharan Africa. *Glob. food sec.* 14:.23-30.

Ma, P., Bian, X., Jia, Z., Guo, X. and Xie, Y., 2016. De novo sequencing and comprehensive analysis of the mutant transcriptome from purple sweet potato (*Ipomoea batatas* L.). *Gene* 575(2): 641-649.

Mcharo, M. and La Bonte, D., 2007. Genotypic variation among sweet potato clones for ß-carotene and sugar content. In *Proceedings of the 13th ISTRC Symposium, Arusha, Tanzania*: 746-754.

Maeshima, M., Sasaki, T. and Asahi, T. 1985. Characterization of major proteins in sweet potato tuberous roots. *Phytochem.* 24(9): 1899-1902.

Magoon, M.L., Krishnan, R. and Bai, K.V., 1970. Cytological evidence on the origin of sweet potato. *Theor and Appl. Genet.* 40(8): 360-366.

Mano H., Ogasawara F., Sato K., Higo H. and Minobe Y. 2007. Isolation of a regulatory gene of anthocyanin biosynthesis in tuberous roots of purple-fleshed sweet potato. *Plant Physiol.* 143: 1252–1268.

Maoka, T., Akimoto, N., Ishiguro, K., Yoshinaga, M. and Yoshimoto, M. 2007. Carotenoids with a 5, 6-dihydro-5, 6-dihydroxy-β-end group, from yellow sweet potato "Benimasari", *Ipomoea batatas* Lam. *Phytochem.* 68(13): 1740-1745.

Marino, R., Ponnaiah, M., Krajewski, P., Frova, C., Gianfranceschi, L., Pè, M.E. and Sari-Gorla, M., 2009. Addressing drought tolerance in maize by transcriptional profiling and mapping. *Mol. Genet. and Genomics* 281(2): 163-179.

Meksem, K. and Kahl, G. 2006. *The handbook of plant genome mapping: genetic and physical mapping*.

Monclus, R., Leplé, J.C., Bastien, C., Bert, P.F., Villar, M., Marron, N., Brignolas, F. and Jorge, V. 2012. Integrating genome annotation and QTL position to identify candidate genes for productivity, architecture and water-use efficiency in Populus spp. *BMC plant biology* 12(1): 173.

Monden, Y., Hara, T., Okada, Y., Jahana, O., Kobayashi, A., Tabuchi, H., Onaga, S. and Tahara, M. 2015. Construction of a linkage map based on retrotransposon insertion polymorphisms in sweet potato via high-throughput sequencing. *Breed. sci.* 65(2): 145-153.

Munoz, P., Carruthers, T., Wood, J.R., Williams, B.R., Weitemier, K., Kronmiller, B., Ellis, D., Anglin, N.L., Longway, L., Harris, S.A. and Rausher, M.D. 2018. Reconciling conflicting phylogenies in the origin of sweet potato and dispersal to Polynesia. *Curr. Biol.* 28(8): 1246-1256.

Mwanga, R.O.M. and Ssemakula, G. 2011. Orange-fleshed sweet potatoes for food. Health and wealth in Uganda. *Int.J. Agric.Sustain.* 9: 42-49.

Nair, A.G., Vidya, P., Ambu, V., Sreekumar, J. and Mohan, C., 2017. Evaluation of orange fleshed sweet potato genotypes for storage root yield and dry matter content. *Skin*3(9).

Niizu, P.Y. and Rodriguez-Amaya, D.B., 2005. New data on the carotenoid composition of raw salad vegetables. *J. Food Comput. Anal.* 18(8): 739-749.

Oki, T., Nagai, S., Yoshinaga, M., Nishiba, Y. and Suda, I. 2006. Contribution of β-carotene to radical scavenging capacity varies among orange-fleshed sweet potato cultivars. *Food sci. and Technol. Res.* 12(2): 156-160.

Paiva, S.A. and Russell, R.M. 1999. β-carotene and other carotenoids as antioxidants. *J. of the Amer. college of nutr.* 18(5): 426-433.

Pflieger, S., Lefebvre, V. and Causse, M. 2001. The candidate gene approach in plant genetics: a review. *Mol. Breed.* 7(4): 275-291.

Quraishi, U.M., Pont, C., Ain, Q.U., Flores, R., Burlot, L., Alaux, M., Quesneville, H. and Salse, J. 2017. Combined genomic and genetic data integration of major agronomical traits in bread wheat (*Triticum aestivum* L.). *Front. in plant sci.* 8: 1843.

Ranjan, P., Yin, T., Zhang, X., Kalluri, U.C., Yang, X., Jawdy, S. and Tuskan, G.A. 2010. Bioinformatics-based identification of candidate genes from QTLs associated with cell wall traits in Populus. *BioEnerg Res.* 3(2): 172-182.

Rao, A.V. and Rao, L.G. 2007. Carotenoids and human health. *Pharm. Res.* 55(3): 207-216.

Rohdich, F., Zepeck, F., Adam, P., Hecht, S., Kaiser, J., Laupitz, R., Grawert, T., Amslinger, S., Eisenreich, W., Bacher, A. and Arigoni, D. 2003. The deoxyxylulose phosphate pathway of isoprenoid biosynthesis: Studies on the mechanisms of the reactions catalyzed by IspG and IspH protein. *Proc. Natl. Acad. Sci. USA* 100: 1586–1591.

Rose, I.M. and Vasanthakaalam, H. 2011. Comparison of the nutrient composition of four sweet potato varieties cultivated in Rwanda. *Am. j. food nutr.* 1(1): 34-38.

Roullier, C., Duputié, A., Wennekes, P., Benoit, L., Bringas, V.M.F., Rossel, G., Tay, D., McKey, D. and Lebot, V. 2013. Disentangling the origins of cultivated sweet potato (*Ipomoea batatas* (L.) Lam.). *PLoS One* 8(5): e62707.

Schafleitner, R., Tincopa, L.R., Palomino, O., Rossel, G., Robles, R.F., Alagon, R., Rivera, C., Quispe, C., Rojas, L., Pacheco, J.A. and Solis, J. 2010. A sweet potato gene index established by de novo assembly of pyrosequencing and Sanger sequences and mining for gene-based microsatellite markers. *BMC genomics* 11(1): 604.

Schweigert, F.J., Steinhagen, B., Raila, J., Siemann, A., Peet, D. and Buscher, U. 2003. Concentrations of carotenoids, retinol and α-tocopherol in plasma and follicular fluid of women undergoing IVF. *Hum. Reproduction*, 18(6): 1259-1264.

Scott, G.J. 2000. Roots and tubers in the global food system: A vision statement to the year 2020*: Report to the technical advisory committee (TAC) of the Consultative Group on International Agricultural Research (CGIAR).*

Shen, Y.H., Yang, F.Y., Lu, B.G., Zhao, W.W., Jiang, T., Feng, L., Chen, X.J. and Ming, R., 2019. Exploring the differential mechanisms of carotenoid biosynthesis in the yellow peel and red flesh of papaya. *BMC genomics* 20(1): 49.

Street, N.R., Skogström, O., Sjödin, A., Tucker, J., Rodríguez- Acosta, M., Nilsson, P., Jansson, S. and Taylor, G. 2006. The genetics and genomics of the drought response in Populus. *The Plant J.*48(3): 321-341.

Sugiura, M. 2015. β-cryptoxanthin and the risk for lifestyle-related disease: findings from recent nutritional epidemiologic studies. *Yakugaku zasshi: J. of the Pharm. Soc. of Japan* 135(1): 67-76.

Takada-Ohara, A., Kumagai, T., Kuranouchi, T., Nakamura, Y., Fujita, T., Nakatani, M., Tamiya, S. and Katayama, K. 2016. 'Aikomachi', a new sweet

potato cultivar with good appearance and high confectionery quality. *Theor. and Appl. Genet*. 30(8): 603-663.

Takahata, Y. 2014. Sweet potato in Japan: Past and future. In *Proceedings of NARO International Symposium (6th Japan-China-Korea Sweet potato Workshop), November*: 28-30.

Takahata, Y., Noda, T. and Nagata, T. 1993. HPLC determination of β-carotene content of sweet potato cultivars and its relationship with color values. *Japanese J. of breed.* 43(3): 421-427.

Tao X., Gu Y.H., Jiang Y.S., Zhang Y.Z. and Wang H.Y. 2013. Transcriptome analysis to identify putative floral-specific genes and flowering regulatory-related genes of sweet potato. *Biosci. Biotechnol. Biochem.*77: 2169–2174.

Tao, X., Gu, Y.H., Wang, H.Y., Zheng, W., Li, X., Zhao, C.W. and Zhang, Y.Z. 2012. Digital gene expression analysis based on integrated de novo transcriptome assembly of sweet potato [*Ipomoea batatas* (L.) Lam.]. *PloS one*, 7(4): e36234.

Tecle, I.Y., Menda, N., Buels, R.M., van der Knaap, E. and Mueller, L.A. 2010. solQTL: a tool for QTL analysis, visualization and linking to genomes at SGN database. *BMC bioinforma.*11(1): 525.

Tewe, O.O., Ojeniyi, F.E. and Abu, O.A. 2003. Sweet potato production utilization, and marketing in Nigeria.

Tian, L., Musetti, V., Kim, J., Magallanes-Lundback, M., and DellaPenna, D. 2004 The *Arabidopsis* LUT1 locus encodes a member of the cytochrome P450 family that is required for carotenoid ε-ring hydroxylation activity. *Proceedings of the National Academy of Sciences of the United States of America.* 101(1):402–407.

Tuan, P.A., Kim, J.K., Lee, J., Park, W.T., Kwon, D.Y., Kim, Y.B., Kim, H.H., Kim, H.R. and Park, S.U., 2012. Analysis of carotenoid accumulation and

expression of carotenoid biosynthesis genes in different organs of Chinese cabbage (Brassica rapa subsp. pekinensis). *EXCLI J.* 11: .508.

Ukoskit, K. and Thompson, P.G. 1997. Autopolyploidy versus allopolyploidy and low-density randomly amplified polymorphic DNA linkage maps of sweet potato. *J. Am. Soc. for Horti. Sci.* 122(6): 822-828.

Van Jaarsveld, P.J., Faber, M., Tanumihardjo, S.A., Nestel, P., Lombard, C.J. and Benadé, A.J.S. 2005. β-Carotene–rich orange-fleshed sweet potato improves the vitamin A status of primary school children assessed with the modified-relative-dose-response test. *The Am. J. Clin Nutr.* 81(5): 1080-1087.

Wang, C.M., Liu, P., Yi, C., Gu, K., Sun, F., Li, L., Lo, L.C., Liu, X., Feng, F., Lin, G. and Cao, S. 2011. A first generation microsatellite-and SNP-based linkage map of Jatropha. *PloS one*, 6(8): e23632.

Wang, Z., Fang, B., Chen, J., Zhang, X., Luo, Z., Huang, L., Chen, X. and Li, Y. 2010. De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (Ipomoea batatas). *BMC genomics* 11(1): 726.

Wang, Z., Li, J., Luo, Z., Huang, L., Chen, X., Fang, B., Li, Y., Chen, J. and Zhang, X. 2011. Characterization and development of EST-derived SSR markers in cultivated sweet potato (Ipomoea batatas). *BMC plant biology*, 11(1): 139.

Wong, J.C., Lambert, R.J., Wurtzel, E.T. and Rocheford, T.R. 2004. QTL and candidate genes phytoene synthase and ζ-carotene desaturase associated with the accumulation of carotenoids in maize. *Theor. Appl. Genet.* 108(2): 349-359.

Woolfe, J.A. 1992. Sweet potato: an untapped food resource. Cambridge University Press.

Wu, S., Lau, K.H., Cao, Q., Hamilton, J.P., Sun, H., Zhou, C., Eserman, L., Gemenet, D.C., Olukolu, B.A., Wang, H. and Crisovan, E. 2018. Genome

sequences of two diploid wild relatives of cultivated sweet potato reveal targets for genetic improvement. *Nat.* (1): 4580.

Xie, F., Burklew, C.E., Yang, Y., Liu, M., Xiao, P., Zhang, B. and Qiu, D. 2012. De novo sequencing and a comprehensive analysis of purple sweet potato (*Ipomoea batatas* L.) transcriptome. *Plant.* 236(1): 101-113.

Xu, E., Vaahtera, L., Hõrak, H., Hincha, D.K., Heyer, A.G. and Brosche, M. 2015. Quantitative trait loci mapping and transcriptome analysis reveal candidate genes regulating the response to ozone in *Arabidopsis thaliana*. *Plant, cell & environ.* 38(7): 1418-1433.

Yamakawa, O. 1999. " Sunny-Red": a new sweet-potato (*Ipomoea batatas*) cultivar for powder. *Bull. Natl. Agric. Res. Cent. Kyushu Okinawa Region* 35: 19-40.

Yamakawa, O., Yoshinaga, M., Kumagai, T., Hidaka, M., Komaki, K., Kukimura, H. and Ishiguro, K. 1998. " J-Red", a new sweet potato cultivar. *Bulletin of the Kyushu National Agricultural Experiment Station (Japan)*.

Yang, J., Moeinzadeh, M.H., Kuhl, H., Helmuth, J., Xiao, P., Liu, G., Zheng, J., Sun, Z., Fan, W., Deng, G. and Wang, H. 2016. The haplotype-resolved genome sequence of hexaploid *Ipomoea batatas* reveals its evolutionary history. *bioRxiv*, : 064428.

Yano, K., Takashi, T., Nagamatsu, S., Kojima, M., Sakakibara, H., Kitano, H., Matsuoka, M. and Aya, K. 2012. Efficacy of microarray profiling data combined with QTL mapping for the identification of a QTL gene controlling the initial growth rate in rice. *Plant and Cell Physiol.* 53(4): 729-739.

Ye, J., Liu, P., Zhu, C., Qu, J., Wang, X., Sun, Y., Sun, F., Jiang, Y., Yue, G. and Wang, C. 2014. Identification of candidate genes JcARF19 and JcIAA9 associated with seed size traits in *Jatropha*. *Funct. integrated genomics*, 14(4): 757-766.

Ye, J., Yang, Y., Chen, B., Shi, J., Luo, M., Zhan, J., Wang, X., Liu, G. and Wang, H. 2017. An integrated analysis of QTL mapping and RNA sequencing provides further insights and promising candidates for pod number variation in rapeseed (*Brassica napus* L.). *BMC genomics,* 18(1): 71.

Yoon, J.B., Kwon, S.W., Ham, T.H., Kim, S., Thomson, M., and Hechanova, S.L. 2015. "Marker-Assisted Breeding", in Current Technologies in Plant Molecular Biology. *Springer:* 95–144.

Yoshinaga, M. 2006. New varieties for dried sweet potato products Hamakomachi and Kyushu No. 137. *Sweet potato Res. Front, 20*: 3.

Yu, B. and Hinchcliffe, M. 2011. *In silico tools for gene discovery*. New York, NY: Humana Press.

Yu, X.X., Ning, Z.H.A.O., Hui, L.I., Qin, J.I.E., Hong, Z.H.A.I., HE, S.Z., Qiang, L.I. and LIU, Q.C. 2014. Identification of QTLs for starch content in sweet potato (Ipomoea batatas (L.) Lam.). *J. of integrated agri.* 13(2): 310-315.

Zhang, D., Cervantes, J., Huamán, Z., Carey, E. and Ghislain, M. 2000. Assessing genetic diversity of sweet potato (*Ipomoea batatas* (L.) Lam.) cultivars from tropical America using AFLP. *Genet. Resour. Crop Evol.* 47(6): 659-665.

Zhang, D., Zhang, H., Chu, S., Li, H., Chi, Y., Triebwasser-Freese, D., Lv, H. and Yu, D. 2017. Integrating QTL mapping and transcriptomics identifies candidate genes underlying QTLs associated with soybean tolerance to low-phosphorus stress. *Plant mol. Biol.* 93(1-2) :137-150.

Zhao, N., Yu, X., Jie, Q., Li, H., Li, H., Hu, J., Zhai, H., He, S. and Liu, Q. 2013. A genetic linkage map based on AFLP and SSR markers and mapping of QTL for dry-matter content in sweet potato. *Mol. Breed.* 32(4): 807-820.

Ziska,L.H., Runion,G.B., Tomecek, M., Priors.A., Torbet,H.A. and Sicher,R. 2009.An evaluation of cassava, sweet potato and field corns potential carbohydrate sources for bioethanol production in Alabama and Maryland. *Biomass Bioenerg.* 33: 1503 – 1508.

APPENDICES

APPENDIX I

CTAB RNA extraction buffer

Tris HCL (pH=8.0)	100mM	
EDTA	25mM	
NaCl	2M	
CTAB	2%	
Mercaptoethanol	2%(v/v)	Freshly prepared
PVP	2% (w/v)	

Prepared in DEPC treated water

APPENDIX II

TBE Buffer (10 X)

Tris base	107 g
Boric acid	55 g
0.5 M EDTA (pH 8.0)	40 mL

Final volume made up to 1000 mL with distilled water and autoclave before use.

APPENDIX III

70% Ethanol

100% Ethanol -70 mLf

Distilled water- 30 mL

INTEGRATION OF QUANTITATIVE TRAIT LOCUS (QTL) FOR TUBER COLOUR VARIATIONS WITH GENOMIC INFORMATION IN SWEET POTATO (*Ipomoea batatas* L.)

By

RESHMA T. K.

(2014-09-118)

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 AGRICULTUREVELLAYANI, THIRUVANANTHAPURAM 695 522 KERALA, INDIA

2019

ABSTRACT

The study entitled Integration of Quantitative Trait Loci (QTL) for tuber colour variations with genomic information in sweet potato (*Ipomoea batatas* L.) was conducted at section of extension and social sciences, ICAR-CTCRI. The main objective of the study was to identify the differentially expressed genes for various tuber colours in sweet potato using RNA sequenced data; to integrate QTL information on tuber colour with genomic information in sweet potato and to validate the identified candidate genes. Sweet potatoes are abundant in compounds of biological effects such as β-carotene, phenolic acids and anthocyanins which gives its unique flesh colours. Here, a comparative transcriptomic analysis was performed to reveal the differentially expressed genes in six sweet potato cultivars with varying flesh colours of white, orange and purple. A total of 22,534, 27,431, 22,590 differentially expressed genes were identified in the pairwise analysis of orange and white, orange and purple and purple and white libraries respectively. Among differentially expressed genes, 5472 were upregulated and 17,062 were downregulated in orange compared to white, 11,670 upregulated genes and 15,761 downregulated genes in orange compared to purple, 7,622 were upregulated and 14,968 were downregulated in purple compared to white. Functional annotation of transcripts associated with the carotenoid biosynthesis pathway revealed the genes involved in the carotenoid biosynthesis pathway.

In the present study, alignment of flanking SSR markers sequences of the QTL controlling β-carotene trait was done with the sweet potato genome assembly showed the position of QTL region on the chromosome. Functional annotation of the identified chromosomal region resulted in the identification of five candidate genes for carotenoid biosynthesis from three QTLs for β-carotene. Transcriptome sequencing and fine mapping of QTL are the efficient ways for discovering novel genes involved in main pathways. The identification of agronomically important genes can be utilized for improvement of sweet potato by the introduction of the genes to commercial sweet potato cultivars and for marker assisted selection.