CHAPTER 1

INTRODUCTION

Vine cacti of the genera *Hylocereus* and *Selenicereus* originating from northern South America, Central America, and Mexico, are currently being grown as new exotic fruit crops. The success is attributed to the fruit quality and characteristics including its attractive colours and shape. The genus *Hylocereus* and *Selenicereus* comprises 16 and 20 species respectively (Tel-Zur *et al.,* 2005 and Le Bellec *et al.,* 2006).

At present, only three species are cultivated on commercials scale including *H. undatus, H. polyrhizus* and *S. megalanthus* in Colombia, Israel, Vietnam and Nicaragua (Tel-Zur *et al.,* 2005). Vietnam is the biggest producer of the *Hylocereus* spp. However, only *H. undatus* and *H. polyrhizus* are cultivated on a small scale in Malaysia. In the market, these two fruits are commonly differentiated based on the peel and pulp colour. *H. polyrhizus* fruit has red peel and pulp whereas *H. undatus* fruit has red peel and white pulp. The red colour is due to the presence of betacyanins, a group of pigments derived from betalains. In the commercial industry, betalains can be used as a natural source of food colouring. Furthermore, it contains anti-oxidant properties for protection against certain oxidative stress-related disorders. The flesh of these fruit has mucilaginous texture and thousands of small soft seed distributed throughout the flesh (Le Bellec *et al.,* 2006).

The plantations are affected by various diseases due to the fact that the *Hylocerues* spp. did not originate locally. This had led to poor production of fruits affecting the small scale farmers in terms of economic such as postharvest losses and the need for pesticide and

herbicide usage. Frequent application of high dose herbicide does not solve the problem and in fact may worsen the problem if the fungus develops resistance to the chemical. Establishing a genomic library could serve as a platform to study the gene interaction as opposed to conventional method of cross breeding to obtain disease resistant hybrid which is often time consuming.

The objective of this study is to construct dragon fruit (*H. undatus*) genomic library through bacteriophage. The procedures involved in genomic library construction include isolation of genomic deoxyribonucleic acid (DNA), generation of DNA fragments for cloning, packaging and transduction. The most crucial component of the entire procedure is the generation of the desired DNA fragments.

CHAPTER 2

LITERATURE REVIEW

2.1 BOTANICAL DESCRIPTION

Hylocereus undatus is a dicotyledonous plant belonging to the Cactaceae family which originated from Latin America. It has 11 pairs of chromosome (2n=22), characterised by triangular stems, branches with aerial roots (Fig 2.1a), large flowers that open at night and can bear fruits that weighs between 200-800g. The fruit has red colour peel and white flesh within which thousands of small black seeds are embedded (Fig 2.1b). The most common cultivation technique is through stem cutting and seed germination was not preferred because plants derived from stem cutting can reach the fruiting stage faster than seed germination which normally takes 3 years to reach fruiting stage (Tel-Zur *et al.*, 2004, 2005, Le Bellec *et al.*, 2006 and Lichtenzveig *et al.*, 2000). Taxonomy of *H. undatus* is as follows:

Kingdom	Plantae
Division	Maglinophyta
Class	Magnoloipsida
Order	Caryophyllales
Family	Cactaceae
Genus	*Hylocereus*
Species	*undatus*

Source: Britton and Rose (1963)

Fig. 2.1a and 2.1b: *H. undatus* plant and the fruit.

2.2 ECONOMIC IMPORTANCE

Attractive colour, shape and the fruit quality were the reasons that drew the farmer's attention to grow the plant as an exotic crop. Among the *Hylocereus* and *Selenicereus* species, only 3 species are commercially cultivated namely *H. undatus*, *H. polyrhizus* and *S. megalanthus* (Tel-Zur *et al.*, 2005 and Le Bellec *et al.*, 2006). *S. megalanthus* fruits are tasty and sweet compared to *Hylocereus* spp. However the appearance is not as attractive. The weight of the fruit in these genera determines the economic value.

Vietnam is the biggest producer of the *H. undatus*. With relative to Vietnam, only small scale plantation of *H. undatus* and *H. polyrhizus* exist in Malaysia. The whole fruit can be consumed fresh or processed as fruit juice. In Malaysia, the annual production and plantation size shows an increasing trend as shown in Figure 2.2.

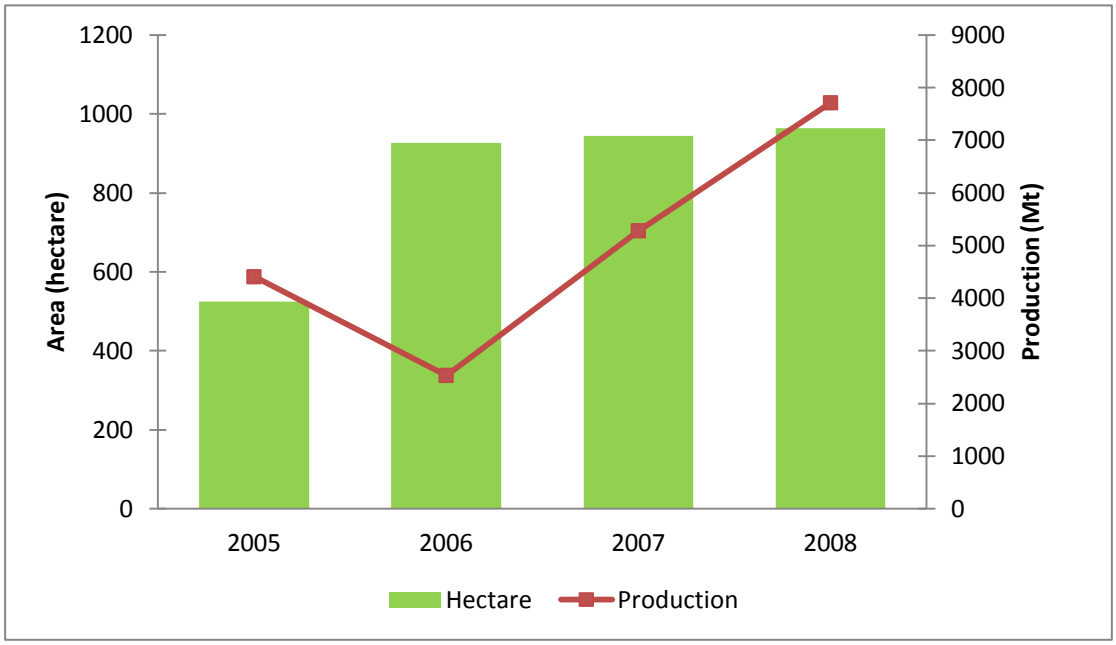

Fig. 2.2: Area of Cultivation and dragon fruit production in Malaysia

Source: www.doa.gov.my

2.3 STUDIES CONDUCTED ON *HYLOCEREUS* SPECIES

There are several studies conducted on *Hylocereus* species whereby most of the studies focus on the physiological part of the fruit. The size of the fruit determines its economic value. From a commercial point of view, it is desirable to have attractive appearance such as size, spinelessness as well as the taste of *S. megalanthus*. In these conducted studies, crosses between the diploid *H. polyrhizus* and the tetraploid *S. megalanthus* as the female and male parent respectively, yielded triploid and aneuploid hybrids producing viable seeds where the number per fruit was strongly dependent on the pollen donor (Tel-Zur *et al.*, 2005).

Besides that, fruit growth, ripening, and the effect of various storage temperatures on the fruit quality of *H. undatus* and *H. polyrhizus* were also studied in Beer-Sheva under

greenhouse conditions. In these studies, it was found that the fruits can retain market quality for a minimum of 2 weeks at 14°C or 1 week at 20°C when harvested near to full colour. This information provided vital information for countries exporting the fruits (Nerd *et al.*, 1999).

There is a growing demand for the use of natural pigments in food colouring to replace synthetic pigments because the natural products are associated with quality and health promotion as opposed to synthetic pigments (Downham and Collins, 2000). Currently, the most important betalain source for natural red colouring is from red beet. However, because of the unfavourable earthlike flavour characteristics caused by geosmin and pyrazine derivatives, as well as high nitrate concentrations associated with the formation of carcinogenic nitrosamines, there is a demand for alternative compounds. Hence, fruits from the Cactaceae such as *H. polyrhizus* have been proposed as a promising betalain source.

2.4 INTRODUCTION TO GENOMIC LIBRARY

There are two types of libraries; genomic and complementary deoxyribonucleic acid (cDNA) library. Genomic library is a population of independent DNA insert containing all the necessary sequence information which represents the total genetic of the organism allowing DNA propagation in the host cell whereas cDNA library are molecules carrying inserts representing messenger ribonucleic acid (mRNA) population from a cell or tissue (Sheffield and Abcouwer). The starting material for the construction of genomic and cDNA library is genomic DNA and mRNA molecules respectively. Both libraries have different applications depending on the objectives of the study. In order to derive the primary sequence of protein or for expression studies cDNA library will then be constructed. If the objective is to study gene interaction or to sequence the structural gene, genomic clone is

then essential. Prior to the construction of cDNA library, identifying the source to obtain the correspondent mRNA is crucial as some mRNAs are predominant in certain types of cell such as globulin in reticulocytes and albumin in hepatocytes. The challenge lies in the construction of cDNA library which is to obtain mRNA molecules that are pure and absent of degradation. As for the construction of genomic library, it requires high quality and intact genomic DNA, manipulation of large insert DNA poses a difficulty such as generating bacterial artificial chromosome (BAC) library. The quality and integrity of the inserts are directly correlated with the success of identifying the gene of interest.

In a nutshell, the construction of the genomic library initially involves the isolation of pure and intact genomic DNA which is followed by the fragmentation of genomic DNA into smaller fragments to be cloned into a chosen vector. In the process of generating recombinant DNA, both the restriction enzymes and ligases are used. Restriction enzyme cleaves the DNA molecule at specific recognition sites by breaking the phosphodisester bond whereas ligase joins two fragments together. In the ligation process, the vector DNA and DNA insert are mixed together so that the complimentary end of the restriction site can be base pair. The complementary sticky ends base pairs by weak hydrogen bonding forming a relatively stable structure for the enzyme. As for the blunt end ligation, the DNA concentration has to be high to increase the chances of association. Ligation takes place between the 5'-P and 3'-OH termini. During ligation, the ligase enzyme catalyses the formation of phosphodiester bond requiring the supply of adenosine triphosphate (ATP) which is contained in the commercial buffer. Alternatively linker, adaptors and homopolymer tailing can be used to produce sticky ends. These recombinant molecules will be then introduced into the host supplying the machinery needed for the replication process to pass on to their progeny (Abcouwer, 2001).

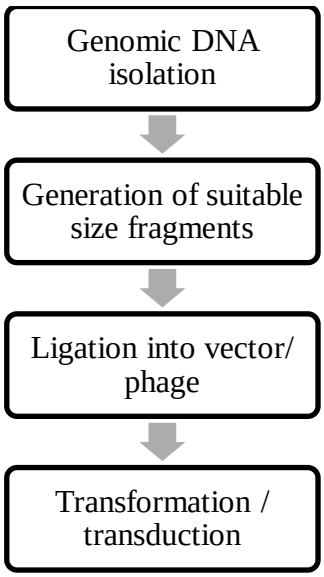

Fig 2.3: The steps in constructing genomic library

Up to date, most genomic libraries are constructed using BAC and yeast artificial chromosome (YAC) vectors. These vectors are preferred due to the ability to carry DNA inserts longer than the largest inserts carried by plasmid and bacteriophage vectors. As the name imply, these vectors masquerade as chromosomes in these cells, being duplicated and distributed equally between daughter cells during synthesis, cytokinesis, and mitosis.

BAC has several advantages with relative to YAC. Firstly, the transformation rate is higher whereby the generation of BAC libraries is faster, easier and smaller amount of DNA are required. Secondly, *E. coli* grows faster and maintenance is easier. Lastly, the plasmid can be isolated using standard miniprep procedures.

The aim of any genome project is to determine the exact sequence of an organism. With this knowledge, information of the complete set of gene function can be obtained. Examples of organisms that had been fully sequenced include *Arabidopsis thaliana*,

Drosophila and *Homo sapiens*. *Arabidopsis thaliana* is the first sequenced plant genome; the genome is 125Mb, densely populated with genes containing approximately 25,500 gene followed by the rice (460Mb). *Arabidopsis thaliana* and rice both represents the dicots and the monocots respectively laying the foundations for future explorations of plant genomics. Both these plants have their own characteristics. *Arabidopsis thaliana* and *Oryza sativa* has the smallest genome in the dicots and the monocots. Comparative approaches can be used when the library is established to transfer the basic knowledge to other species (Klaus and Hans-Werner, 2001).

The other hallmark genome project was the Human Genome Project (HGP), which cost US$3.0 billion started on October 1, 1990 and completed on September 30, 2005. This project involved 32 scientists from 14 countries mainly from United State, United Kingdom, France and Japan. With this library, they were able to detect subtle genetic variations which made people susceptible to cancer and heart disease. Furthermore, the map of genetic markers will be an invaluable source for the location of disease genes. Some of the findings in the project were as follows: (1) Only 30,000-40,000 protein coding genes in 3 billion nucleotides; (2) Mutation rate in males are twice higher than female in meiosis; (3) Genes are more complex with more alternative splicing generating a larger number of protein products (Falcon, 2002).

Due to the rapid development of sequencing technology and cost reduction, more organisms can be sequenced. However, sequence of large plant genome which greatly exceeds human genome cannot be anticipated in the near future as it requires huge financial resources. Examples of plant genomes significantly larger than human genome include bread wheat (17,000 Mb) and oat (26,000Mb). Therefore, before the initiation of a genome

project, factors such as cost, feasibility and economic importance of the subject have to be taken into consideration (Klaus and Hans-Werner, 2001).

To construct a genomic library, a few factors have to be taken into consideration which ultimately depends on the purpose of the study or the size of the genome. If the objective of the study involves physical mapping or genome sequencing, then it requires the isolation of protoplast, manipulation of large insert and BAC vector usage. However, if there is prior knowledge to the gene of interest with falls in the size range smaller, lambda vector can be used and isolation of genomic DNA would be easier. Besides that, bacteriophage vector is used when the library are expected to be used frequently. Cloning vector with high capacity significantly reduces the number of clones hence less time required to conduct screening in the later stage.

For instance, *E. coli* with genome size of 4.6×10^6 base pairs is fragmented into 17kb fragments length where 820 clones is needed to find the particular gene of interest compared with a 35kb length if fragments which will produce 410 clones with a 95% probability. This is calculated based on this formula: $N = \ln(1-P) / \ln[1-(a/b)]$ where N, is the number of clones needed; P, is the probability of that any gene will be present; a, is the average size of the DNA fragments inserted into the vector and b, is the total size of the genome. This equation derived from Clarke & Carbon (1976) was used to calculate for each library the probability to and any given DNA sequence represented in two libraries which is determined by the number of recombinants, the size of the genome and the average insert size (Brown, 2001 and Wang *et al.*, 2003).

Due to lack of information pertaining to *Hylocereus undatus* in the aspect of molecular work and problem associated of it affected with disease, bacteriophage genomic library approaches was employed because of its relative simplicity in construction compared to the BAC library. Lambda (λ) serves a good candidature as a cloning vector due to several reasons. One obvious reason is that it can accommodate bigger fragment of DNA compared to plasmid. The genome of λ-phage is a double-stranded DNA molecule approximately 50kb in length and central third of the viral genome is not essential for lytic growth and, thus, can be replaced by a variety of foreign gene segments. λ-vectors usually contain multiple cloning sites facilitating the cloning of foreign DNA. In addition, some of the regulatory regions are dispersed away from the structural gene. The efficiency of DNA transformation also is higher due to its natural ability of injecting DNA into the host cells. Commonly, λ-phage vectors replicate via the lytic pathway. During lytic growth, the viral DNA is replicated manifold, a large number of phage gene products are synthesized, and progeny phage particles are assembled. The cell is eventually lysed, releasing its many new infectious virus particles; at this time, plaques will form on an infected bacterial lawn. In addition, λ-phage vectors are suitable for screening using nucleic acid probes. In terms of gene screening, genomic DNA libraries are often screened by hybridization using a radioactive nucleic acid probe (Wang *et al.,* 2003). After establishing genomic library, each individual clone can be isolated out for sequencing. Therefore in any sequencing project, library needs to be constructed first. It will provide opportunity for defining targets for manipulation to achieve disease resistance by designing the resistance gene that recognise key components of pathogens which will induced defence response (Richard Michelmore, 2000).

There are few strategies that can be used to cloned resistance gene such as tranposon tagging, map base cloning, cross hybridisation and polymerase chain reaction (PCR) methods. As more resistance gene are being cloned, PCR based method can be employ to isolate even more resistance gene on the basis of sequence similarity or conserved domains. However, most of the different resistance genes have different specifications hence region of sequence divergent will occur and thus will provide an opportunity for the development of gene specific PCR.

2.5 EXTRACTION, PURIFICATION AND QUANTIFICATION OF DNA

The prerequisite of constructing genomic library was to isolate of pure and intact genomic DNA from *Hylocereus undatus*. Isolation of high quality nucleic acid from plants is difficult due to the presence of polysaccharides and high polyphenol compounds in the plant cells such as cactus and lychee (Tel-Zur *et al.*, 1999 and Puchooa, 2004). Plant cells have cell wall that is not easily removed with physical mean without damaging the cell contents. The vacuole which is full of degradative enzymes and secondary metabolites such as DNase and phenolic compounds with can damage the DNA. In any DNA extraction protocols, the ultimate aim is to isolate high quality and intact DNA. The protocol should be quick and simple which can produce adequate yield for downstream process such as polymerase chain amplification and restriction digest. Avoiding the usage of dangerous chemical such as phenol is even better (Puchooa, 2004).

Furthermore, the chemical composition in plant cell differs according to the parts of the plant. This was reported in Tel-Zur *et al.*, (1999) where they used roots as starting materials due to the lower viscosity of the extracts relative to that other of other plant tissues. Besides that, the chemical profiles are also influenced by time such as mature grapevine leaves

contains high quantities of polysaccharides, polyphenols and tannin when no fresh or actively growing young leaves were available (Hanania *et al.*,2004). Photosynthetic active tissue contains phenolic compounds that oxidize during extraction will irreversibly bind with nucleic acid to form gelatinous matrix (Michiels *et al.*, 2003). Therefore, proper choice of leaf tissue is very crucial in DNA extraction. Young and actively dividing tissues such as non-fully expanded leaf is preferred as it contains high cell mass and does not contains secondary compounds (Diadema *et al.*, 2003).

Tissues extraction with cetylmethylammonium bromide (CTAB) in high salt condition effectively removed polysaccharide by increasing their solubility in ethanaol therefore suppressed the co-precipitation of polysaccharides and DNA, the basis of CTAB extraction method. To overcome the interference caused by secondary metabolites, chemical such as polyvinylpyrrolidones (PVP), β-mercaptoethanol, and bisulphite were often added into the extraction buffer (Murray and Thompson, 1980, Micheils *et al.*, 2003, De la Cruz *et al.*, 1997, Maliyakal, 1992, Salzman *et al.*, 1999 and Lodhi *et al.*, 1994).

Once the genomic DNA was obtained, the next step was to digest genomic DNA to produce clonable fragments. Genomic DNA contains many repetitive elements in which the distribution of the cleavage sites is no longer random so that some of the many restriction enzymes present in this smear produce clearly defined, typical bands.

Purity of nuclear DNA has a significant impact on the degree of digestion and the quality of library. By the addition of washing steps, high quality and stable high molecular weight DNA can be obtained. The additional washing steps in DNA extraction also resulted in a low percentage of chloroplast or mitochondrial DNA contamination. In order to increase

the quality and yield of partially digested DNA, different strategies can be employed in the experiments: aliquot of genomic DNA can be digested in several tubes rather than one single tube alone to increase the surface/volume ratio in order to improve access of the restriction enzyme to the high molecular weight DNA (Rahman *et al.*, 2007). The size-fractionated DNA fragments were recovered by electroelution rather than by the common method of melting gel slices followed by agarose treatment. By this method, the yielded DNAs were comparatively less degraded and more amenable to ligation. In addition, a further advantage of the electroelution method is the ability to use regular high melting agarose, which makes gel handling easier than the low melting agarose required for enzyme-based DNA extraction. Tris-acetate-EDTA (TAE) buffer systems have been employed in the electrophoretic separation of partially digested DNA, as borate ions may inhibit ligation reaction used in the construction of BAC libraries (He *et al.*, 2003)

2.6 VALIDATE THE LIBRARY (GENOME COVERAGE)

Once the library was successfully constructed, the library has to be validated to verify whether it is representative of the genome of the organism. There are few approaches to do so; one common approach will be using the Clarke & Carbon equation: $N= \ln (1-P) / \ln [1-(a/b)]$ where N, is the number of clones needed; P, is the probability of that any gene will be present; a, is the average size of the DNA fragments inserted into the vector and b, is the total size of the genome. However this approach only feasible provided there is information regarding the genome size. It calculates the probability of any given DNA sequence in the library. The average size of insert DNA was obtained by randomly selecting the clones and digest with specific restriction enzyme to release the insert DNA and analyse in gel electrophoresis. There are various method that can be used to estimate the genome size. Nuclear DNA staining methods, reassociation kinetics, genomic reconstruction, and direct

contour-clamped homogeneous electric field (CHEF) gel analysis of chromosome numbers and sizes was used to determine the fungal and oomycete species. However, these methods have led to inconsistent estimations of the genome size in some species, particularly those whose genomes are large and contain chromosomes that in CHEF gels are larger than the size markers and difficult to resolve. Example of it, the genome size of *Phytophthora sojae* was estimated to be 46.5Mb from analysis of CHEF gel electrophoresis, which was about 50% of the value estimated from nuclear DNA staining methods and 75% of that estimated by using DNA reassociation kinetics and genomic reconstructions (Shan and Hardham, 2004). Alternatively colony hybridization can be employ to determine the genome coverage by hybridizing with probe generated from single copy gene or know target sequence which can be obtained from database (Rahman *et al.*, 2007).

Furthermore clone generated from the chloroplast and mitochondria DNA also have to be subtracted as this is not genomic DNA. Chloroplast or mitochondria specific gene can be used to screen and estimate the contamination (Lin *et al.*, 2006).