J. Trop. Agric. and Fd. Sc. 44(2)(2016): 219 – 228

Sequence information on single nucleotide polymorphism (SNP) through genome sequencing analysis of *Carica papaya* variety Eksotika and Sekaki

Z.A. Rabiatul Adawiah[1], A.B. Norliza[2], Y. Muhammad Fairuz[1], A. Norzihan[1] and A.B. Umi Kalsom[3]

[1]Technical Service and Lab Centre, MARDI Headquarters, Persiaran MARDI-UPM, 43400 Serdang, Selangor, Malaysia
[2]Biotechnology and Nanotechnology Research Centre, MARDI Headquarters, Persiaran MARDI-UPM, 43400 Serdang, Selangor, Malaysia
[3]Director General Office, MARDI Headquarters, Persiaran MARDI-UPM, 43400 Serdang, Selangor, Malaysia

Abstract
Carica papaya L. is a tropical fruit and an important crop grown for export and local consumption. In this study, three papaya varieties (Solo, Eksotika and Sekaki) genome were sequenced using next generation sequencing platform for single nucleotide polymorphism (SNP) identification and selection. The assembly of Solo sequence reads and Sun Up papaya draft genome were used as reference sequences to align the two papaya genomes (Eksotika and Sekaki). The assembly of Solo sequence reads and Sun Up contigs have generated 23,318 scaffolds with the length of scaffolds above 200 bp. The papaya SNP discovery yielded 934 and 7,959 putative SNPs in Eksotika and Sekaki, respectively. Our study demonstrated the utilisation of next generation sequencing technology coupled with bioinformatics approach for the selection of putative SNP markers that will be valuable for the development of papaya SNP markers. The SNPs identified will enable high throughput genotyping of papaya germplasm as well as a tool for marker assisted selection in papaya breeding program.

Keywords: *Carica papaya,* genome sequencing, *in silico*, molecular marker, next generation sequencing, single nucleotide polymorphism

Introduction

Carica papaya L. is a tropical fruit native to Southern Mexico and Central America. It has been widely cultivated in all parts of the world including South East Asian region. According to the United States Recommended Daily Allowance (USRDA), papaya is ranked first for its high nutritional quality and medical properties. In comparison to 38 other common fruits, papaya is highly recommended in daily diet due to the presence of vitamin A, vitamin C, potassium, folate, niacin, thiamine, riboflavin, iron, calcium and fibre in the fruits. The fruit, stems, leaves and roots also offer a wide range of medical properties including papain, a valuable proteolytic enzyme (Thomas et al. 2009).

The global production of papaya fruits has been increasing due to popular demands from consumers, resulting in papaya as the third ranking crop among tropical fruits behind mango and pineapple. In Malaysia, *Carica papaya* is an important crop with an estimated export value of RM100 – RM120 million per year (Rabu and Mat Lin 2005).

Article history
Received: 1.7.15
Accepted: 3.2.16

Authors' full names: Rabiatul Adawiah Zainal Abidin, Norliza Abu Bakar, Muhammad Fairuz Yusof, Norzihan Abdullah and Umi Kalsom Abu Bakar
E-mail: rabiatul@mardi.gov.my

Eksotika and Sekaki are among the important papaya cultivars in Malaysia. Sekaki with medium size fruit (1.5 – 2.0 kg) is mainly grown for domestic consumption while Eksotika forms the backbone of the papaya export industry (Chan 1997).

However, due to the outbreak of bacterial dieback disease and fruit fly quarantine restrictions from China, papaya production in Malaysia has decreased by 60% since 2005 (Chan and Baharuddin 2010). Therefore, there is an urgent need to revive the Malaysia papaya industry through breeding and utilising genomics approaches.

Research in papaya 'omics' has increased progressively in the past years. The information based on genomics, genetics and bioinformatics research is useful for papaya researchers. Through the advancement of high throughput sequencing and molecular marker technology, it is now possible to assist plant breeding programme by identifying single nucleotide polymorphism (SNP) markers from plant genomes (Fukushima et al. 2009). Still, this information has not been fully utilised for the improvement of papaya breeding in Malaysia.

SNP is often used as a genetic marker of choice due to its abundance and wide applications (Duran et al. 2009). The set of SNP markers will be useful in constructing genetic map, determining association between genotype and phenotype as well as aligning physical and genetic maps (Duran et al. 2009).

Recently, the SNP discovery from whole genome sequences has been adopted by economically important temperate fruit crops such as peach (*Prunus*) (Ahmad et al. 2011), sweet cherry (*Prunus avium*) (Peace et al. 2012), pear (*Pyrus*) (Bartlett et al. 2014), apple (*Malus*) (Bianco et al. 2014) and strawberry (*Fragaria*) (Bassil et al. 2015) as a genetic marker tool to facilitate more precise selection of breeding populations in fruit cultivars. As a result, a total of 9K, 6K, 8K, 20K and 90K SNP array were developed from peach, sweet cherry, pear, apple and strawberry, respectively using genotyping technology (Ahmad et al. 2011; Peace et al. 2012; Bartlett et al. 2014; Bianco et al. 2014; Bassil et al. 2015). This suggested that the papaya SNP markers could be developed using NGS and genotyping technology.

On the other hand, the discovery of SNP markers from whole genome sequences of tropical fruits is less developed as compared to temperate fruit crops. There are limited published reports on the development of papaya SNP markers from genome sequencing except for a recent expressed sequence tags analysis that discovered a certain number of papaya SNPs (Zeng et al. 2014). Nevertheless, the ESTs approach is limited to specific expressed regions of plant genome and is restricted either to certain tissue type or environmental condition. In addition to this, the papaya EST sequences were also from a limited number of papaya cultivars.

To further adopt the advent of the next generation sequencing technology in papaya research, we have embarked on the whole genome sequencing of three important papaya varieties namely Solo, Eksotika and Sekaki. Solo is the parent of Eksotika, whereas Eksotika and Sekaki are the two most widely grown papaya in Malaysia. Despite being susceptible to dieback disease, Eksotika has other excellent quality traits. Eksotika has a small fruit size (400 – 800 g) with high sugar content of 12 – 14 °Brix and orange-red flesh which gives off a pleasant aroma. However, the fruit has a short shelf life due to its soft texture and high sensitivity to environmental stress. On the other hand, Sekaki is a medium sized fruit (1.5 – 2.0 kg), flesh red cultivar, coloured and freckle free skin with lower sugar content than Eksotika.

Here, we described how we coupled resequencing of papaya genome and reference based assembly approaches to

discover over 8000 putative SNPs in papaya. This study aimed to identify putative SNP markers obtained from the genome sequences of three papaya varieties. This result will support the development of papaya SNP panel as a genetic tool for papaya trait improvement via linkage analysis, fingerprinting, quantitative trait loci (QTL) and marker assisted selection (MAS).

Materials and methods

Plant material

Papaya cultivar seeds were supplied and grown by breeder (Dr Johari Sarip) from Horticulture Research Centre (HRC), MARDI. All plants were grown in a greenhouse at HRC, MARDI under ambient light and temperature conditions and watered as needed. For sample collection, papaya leaves were collected using sterilised blade and washed with distilled water to remove dirt. Leave samples were stored at –80 °C until further use. *Table 1* shows the list of papaya cultivars used in this study.

Table 1. List of papaya cultivars used in this study

Cultivar	Purpose
Sun Up	Reference genome
Solo	Reference genome
Eksotika	SNP discovery
Sekaki	SNP discovery

Deoxyribonucleic acid (DNA) isolation and whole genome sequencing

Leaves from four-month old Solo, Eksotika and Sekaki papaya were harvested and grounded in liquid nitrogen. About 200 mg of the grounded samples were aliquoted in 2 ml microcentrifuge tubes and labelled prior total DNA extraction. The genomic DNA prepared from each cultivar was extracted using the method described by Murray and Thompson (1980). The quality and quantity of DNA were analysed using NanoDrop 2000 (Thermo Fisher Scientific Inc, USA) and Qubit 2.0 DNA Broad Range Assay (Invitrogen, USA). The integrity of DNA samples was checked on 1% agarose gel. The extracted DNA was treated with RNase and then subjected to sequencing library preparation according to the manufacturer's protocol, Illumina HiSeq 2000 sequencing (Illumina, Inc, Sand Diego, CA, USA). The resulting library was sequenced for 202 cycles to generate a total of 40 GB raw sequence reads with the length of 101bp for each sequence reads.

Data preprocessing and generation of reference sequences

The raw sequence reads from Solo, Eksotika and Sekaki were pre-processed to generate high quality sequence reads for further analysis steps. FastQC version 0.10.1 (Andrew 2010) was used to evaluate the read quality before and after pre-processing. Trimmomatic version 0.27 (Bolger et al. 2014) was exploited to trim the sequence reads and adapter clips. Prinseq version 0.2.1 (Schmieder and Edwards 2011) was used to remove sequences with ambiguous bases along with duplicated reads.

To generate a reference sequence for SNP discovery, sequence reads of Solo were assembled with draft genome sequence of Sun Up papaya version 3.1 (Ming et al. 2009). We retrieved the Sun Up draft genome from Phytozome database (Goodstein et al. 2012) with a total of 34,2680,090 nucleotide bp/342.68 Mb.

The assembly of reference sequences was carried out using MIRA version 4.0 to construct the contigs (Chevreux et al. 2014), whereas SSPACE (Marten et al. 2010) was utilised to generate papaya scaffolds.

SNP discovery and SNP selection

Eksotika and Sekaki sequence reads were aligned separately to papaya reference sequences using BWA version 0.6.2 (Li et al. 2009). BWA is aligner software that suitable to align DNA sequences against large reference genome.

The resulting sequence alignment of Eksotika and Sekaki in BAM format were then subjected to SNP discovery using SAMTools mpile up version 0.1.18 (Li et al. 2009) with specific parameter(samtools.pl varFilter -d 10, -D 100) retaining SNPs with a minimum read depth of 10, and a maximum read depth of 100. Further filtering of the SNPs was done with a custom Perl scripts that remove SNPs with quality score <100, SNPs without annotated genes, SNPs with another SNPs within 150 bp region and SNPs with insufficient flanking bases. The genomic positions of SNPs and associated SNPs information were stored in a MySQL relational database using BioMart version 0.7 (Damian et al. 2009) applications. *Figure 1* shows the computational analysis pipeline for papaya SNPs discovery.

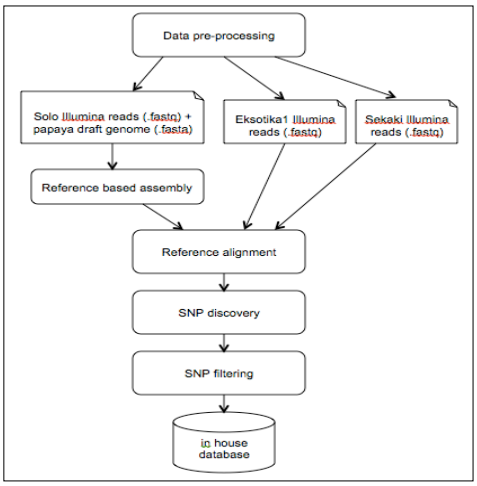

Figure 1. Computational analysis pipeline for discovery of SNP in papaya

Results and discussion

Papaya whole genome sequencing by NGS and data pre-processing

More than 1.6 million bp, 1.49 million bp, 1.5 million bp were sequenced with sequencing coverage of 33x, 31x and 30x for Solo, Eksotika and Sekaki, respectively. This sequencing coverage was estimated based on papaya draft genome size (372 Mb). It was expected that the sequencing coverage was high and suitable for finding SNPs in polymorphic regions (Brockman et al. 2008). *Table 2* shows the summary of next generation sequencing used for SNP discovery in papaya. The length of sequence reads at 101bp is suitable for SNP discovery (Brockman et al. 2008).

Generation of papaya reference sequences and mapping of Eksotika and Sekaki to reference sequences

We performed reference-based assembly of Solo and Sun Up to generate papaya reference sequences. From the assembly, a total of 23,138 scaffolds were generated with 34.86% GC contents. We only kept length of scaffolds that are more than 200 bp. The largest scaffold of reference sequences were 270, 051 bp. The total length of reference sequences was 299.23 Mb which covered 80.44% of the papaya genome size. This has provided an estimated coverage of papaya genome size, assuming the papaya genome size is 372 Mb. *Table 3* shows the statistics of assembly of papaya reference sequences.

To assess the assembly of reference sequences, we aligned the Solo sequence reads against the reference sequences. From the alignment results, 98% of Solo sequence reads were successfully aligned against reference sequences.

Using BWA, the quality sequence reads of Eksotika and Sekaki were successfully mapped to papaya reference sequences. Both alignments produced more than 90% sequence reads that were successfully mapped onto papaya reference sequences. *Table 4* shows the summary of alignment statistics for Eksotika and Sekaki against papaya reference sequences.

Table 2. Summary of sequence reads generated using Illumina HiSeq 2000 sequencing platform

	Solo	Eksotika	Sekaki
Raw data			
Total sequence reads produced (bp)	161, 663, 642	149, 741, 086	156, 058, 148
Sequencing depth	33x	31x	30x
After data preprocessing			
Total sequence reads (bp)	124, 668, 561	118, 277, 310	114, 303, 915

Table 3. Statistics of papaya reference sequences

	Scaffolds
Total scaffolds	23, 318
Total nucleotide size (bp)	299, 238, 687
Longest scaffolds (bp)	270, 051
N50 length (bp)	35, 929
GC content %	34.86

Table 4. Statistics of papaya reference alignment and SNP discovery in Eksotika and Sekaki

Cultivar	Eksotika	Sekaki
Total reads	118, 277, 310	114, 303, 915
Total mapped reads	105, 285,114	100, 503, 060
Total mapped reads (%)	95.77%	96.51%
Total SNPs discovered	495, 832	610, 483
Total SNPs after SNP filtering	934	7959

Analysis of SNP discovery and SNP selection

The SNPs filtering criteria were applied to decrease number of false positive from the identified SNPs as well as to obtain high confident SNPs for SNP validation step. For instance, removing SNPs with insufficient flanking bases is selected as a criteria for SNP filtering to design high quality primers for SNP validation. To focus on the SNPs differences between these two varieties, a total of 1571 SNPs with redundant position between Eksotika and Sekaki were identified. From the results, a total of 934 SNPs were discovered in Eksotika, whereas a total of 7959 SNPs were discovered in Sekaki. *Figure 2* shows the number of SNPs identified between Eksotika and Sekaki as well as the overlapping SNPs.

The analysis also indicated scaffold809 in Eksotika has the highest SNP distribution with 23 total SNPs. The average distances between adjacent SNPs on scaffold809 is 5451bp. The physical map of Sekaki, scaffold91, scaffold798 and scaffold1442 showed the highest SNP distribution with 26

total SNPs (*Figure 3*). The average distance between adjacent SNPs on scaffold798 is 4029.73bp, scaffold91 is 2819.92bp while scaffold1442 is 1179.07bp. The average distance between adjacent SNPs on Eksotika and Sekaki papaya scaffold contributes useful information to construct saturated physical and genetic maps of targeted trait such as disease resistance trait. It has been a focal point for the discovery of SNPs from fruit crops to obtain the high number and well distributed SNPs across the genome to permit the high resolution of physical and genetic maps (Ahmad et al. 2011; Bianco et al. 2014; Bassil et al. 2015).

We also conducted functional annotation of selected SNPs using BLASTX with cut off e-value of $1e\text{-}0^6$ in Eksotika and Sekaki. The selected papaya genes and coding sequences (CDS) were retrieved from Phytozome database. The Phytozome database is a central hub of plant genome information that helps scientific community in analysing their plant genome of interest (Goodstein et al. 2012). Of the 934 total SNPs in Eksotika, 866 SNPs had significant hits to papaya CDS from Phytozome database with matches to 406 known genes. A total of 7200 SNPs from Sekaki had significant BLASTX hits to the Phytozome database and were matched to 2108 known genes (*Table 5*).The BLASTX results were categorised into three categories that are, known genes, putative genes and unknown genes. The functional annotation results from BLASTX analysis are valuable to support the predicted functional SNPs if the SNPs were located in gene coding region that could be needed in the analysis of papaya QTL markers or characterisation of functional genes. *Figure 4* shows the distribution of BLASTX results into three categories in Eksotika and Sekaki.

The following characteristics of putative SNPs information were stored in an in house database using Biomart version 0.7 (Damian et al. 2009).

-scaffold name
-SNP ID
-SNP position
-SNP alel
-2 by 75 bp of flanking position

All the SNPs information is required when designing the primers for SNP detection using genotyping platform. *Figure 5* shows the user interface of the in house papaya SNP database using Biomart application.

One of the limitations in discovering papaya SNPs is that it is dependent on the papaya reference genome. To-date papaya reference genome is available as scaffolds and the genes are not mapped onto the nine chromosomes (Ming et al. 2008). Thus, this disadvantage could lead to the erroneous of SNPs position across the papaya genome. To overcome this limitation, high quality resequencing of various papaya accession would be useful to establish the papaya reference genome prior to carrying out genome wide SNP discovery. This approach can also be used to set up a sophisticated bioinformatics pipeline that can also facilitate the discovery of SNPs in a concise and rapid manner.

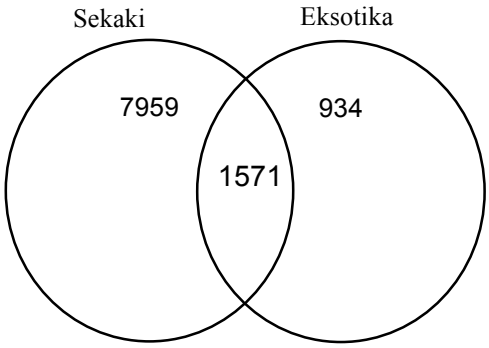

Figure 2.The number of SNPs identified in Eksotika and Sekakias well as the overlapping SNPs

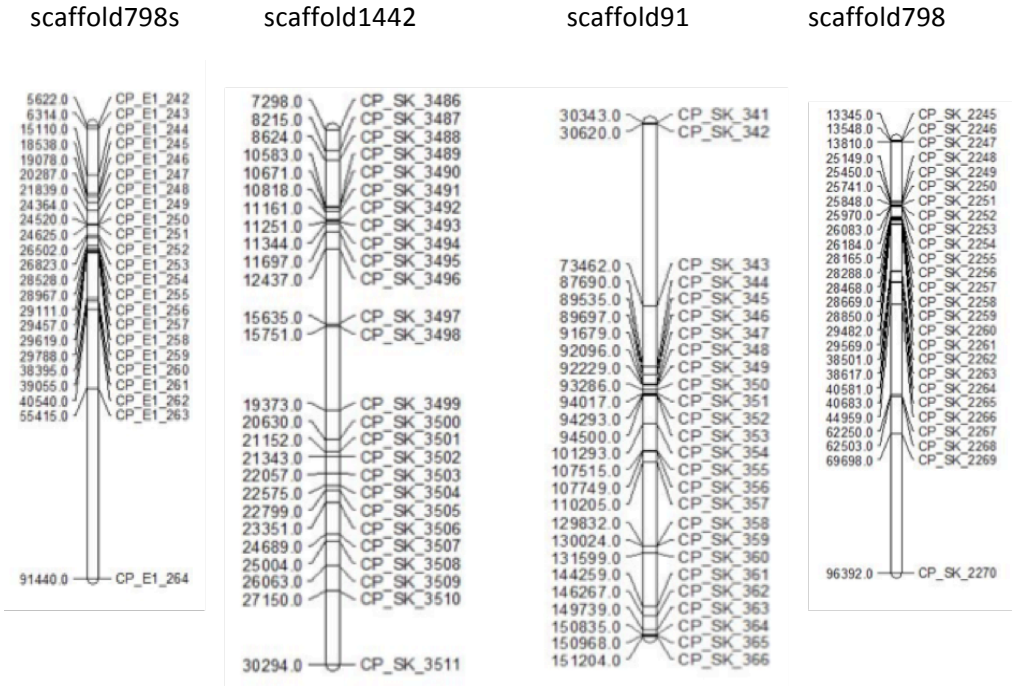

Figure 3. Physical maps of putative SNP markers from papaya scaffold809 in Eksotika, whereas scaffold798, scaffold91 and scaffold1442 in Sekaki.

Table 5. Summary of BLASTX analysis

Variety	Total SNP analysed	Total SNP with significant match to Phytozome database	Total SNP with significant match to known genes
Eksotika	934	866	406
Sekaki	7959	7200	2108
Total	8893	8066	2514

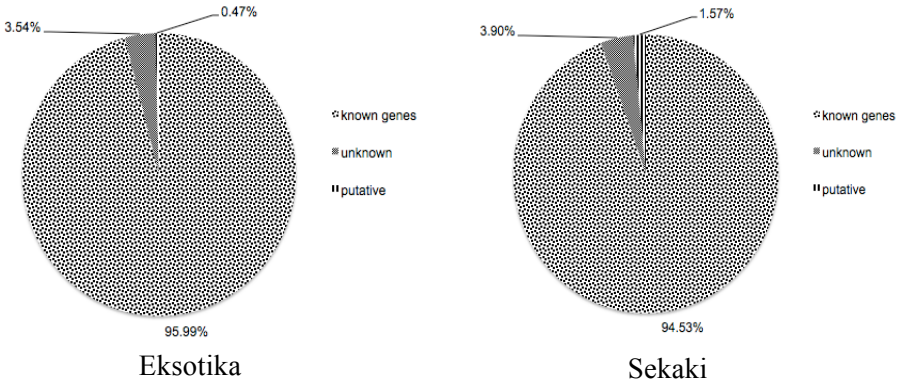

Figure 4. Percentage of SNPs that matched into known genes, unknown genes and putative genes in Eksotika and Sekaki

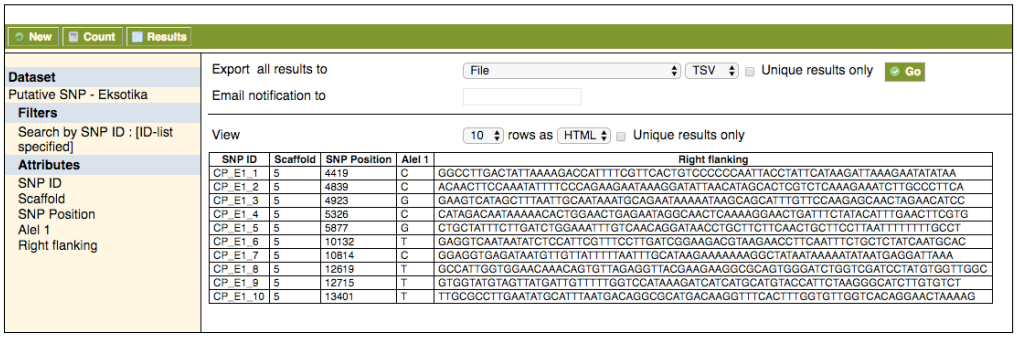

Figure 5. In house papaya SNP database

Conclusion

In summary, by combining next generation sequencing technology of Illumina and bioinformatics approach, we were able to identify an abundance of SNPs in selected papaya varieties. These potential SNPs will be valuable to design a SNP array or SNP assay using genotyping platform such as Mass ARRAY platform (Agena Biosciences, Inc) and KASP, respectively. The papaya SNP panel will provide a useful tool for finger printing study, linkage analysis and QTL mapping in papaya. Additional SNP discovery using the approaches described in this paper is essential in developing optimized SNP panel for identification of suitable elite parental lines.

Acknowledgement

We are thankful to Malaysia Genomic Research Centre (MGRC) team for carrying out the Illumina HiSeq 2000 sequencing and sequence assembly. This research is supported by MARDI and NKEA EPP14.

References

Ahmad, R., Parfitt, D.E., Fass, J., Ogundiwin, E., Dhingra, A., Gradziel, T.M. et al. (2011). Whole genome sequencing of peach (*Prunus persica* L.) for SNP identification and selection. *BMC Genomics* 12(1): 569

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Retrieved on 27 Feb. 2012 from http://www.bioinformatics.babraham.ac.uk

Bartlett, L., Ireland, H., Fiers, M., Dzierzon, H., Cestaro, A., Fontana, P. et al. (2014). The draft genome sequence of european pear (*Pyrus communis* L.'Bartlett'). *PLOS One* 9(4): 1 – 12

Bassil, N.V, Davis,T.M., Zhang, H., Ficklin, S., Mittmann, M., Webster, T. et al. (2015). Development and preliminary evaluation of a 90 K Axiom ® SNParray for the allo-octoploid cultivated strawberry Fragaria × ananassa. *PLOS One* 16(155): 1 – 30

Bianco, L., Cestaro, A., Sargent, D.J., Banchi, E., Derdak, S., Di, M. et al. (2014). Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (*Malus x domestica Borkh). PLOS One* 9(10): 1 – 9

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W.L., Russ, C., Lander, E.S., Nusbaum, C. and Jaffe, D.B. (2008). Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research* 18: 763 – 770

Bolger, A., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 14 – 20

Chan, Y.K. (1997). Backcross method in improvement of papaya (*Carica papaya* L.). *Malaysia Applied Biology* 16: 95 – 100

Chan, Y.K. and Baharuddin, A.G. (2010). Rejuvenating the flagging papaya industry in Malaysia: the role of MAFC. *Acta Horticulture (ISHS)* 851: 37 – 40

Chevreux, M.A. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research* 14: 1147 – 1159

Damian, S., Syed, H., Benoit, B., Richard, H., Darin, L., Gudmundur, T. and Arek, K. (2009). BioMart – biological queries made easy. *BMC Genomics* 10(22): 1 – 12

Duran, C., Edwards, D. and Batley, J. (2009). Molecular marker discovery and genetic map visualization. *In: Bioinformatics: Tools and applications*, (Edwards, D., Stajich, J. and Hansen, D., eds.), p. 165 – 189. New York: Springer

Fukushima, A., Kusano, M., Redestig, H., Arita, M. and Saito, K. (2009). Integrated omics approaches in plant systems biology. *Current Opinion Chemical Biol*ogy 13: 532 – 538

Goodstein, D.M., Shu, S., Howson, R., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* 40: 1178 – 1186

Lehnert, R. (2013). Marked for progress-key genes labeled "jewels in the genome" set the stage for faster, cheaper plant breeding. *Bioinformatics* 25: 54 – 60

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009). 1000 genome project data processing subgroup. The sequence alignment/map (SAM) format and SAM tools. *Bioinformatics* 25: 8 – 9

Marten, B., Christiaan, V.H., Hans, J.J., Derek, B., and Walter P. (2010). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27: 578 – 579

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A. and Saw, J.H. et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* 452: 991 – 997

Murray, H.G. and Thompson, W.F. (1980). Rapid isolation of high molecular weight DNA. *Nucleic Acids Research* 8: 4321 – 4325

Peace, C., Bassil, N., Main, D., Ficklin, S., Rosyara, U. R., Stegmeir, T. et al. (2012). Development and evaluation of a Genome-Wide 6K SNP array for diploid sweet cherry and tetraploid sour cherry. *PLOS One* 7(12): 1 – 15

Rabu, M.R. and Mat Lin, R.(2005). Prospect of papaya in the world market: Malaysia perspective. *Proceeding of first international symposium on papaya*, 22 – 24 Nov. 2005, Genting Higlands, (Chan, Y.K., Paul, Robert, E., Shukor A.R., eds.), Serdang: MARDI

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27: 863 – 864

Thomas, G.E., Rodolfoa, H.G., Juana, M.D., Georginaa, S.F., Luisb, C.G., Ingrida, R.B. and Santiagoa, G.T. (2009). Proteolytic activity in enzymatic extracts from Carica papaya L. cv. Maradol harvest by-products. *Process Biochemistry* 44: 77 – 82

Zeng, F., Yu, Q., Hou, S., Moore, P.H., Alam, M. and Ming, R. (2014). Features of transcriptomein trioecious papaya revealed bya large-scale sequencing of ESTs and comparative analysis in higher plants. *Plant Omics Journal* 7(6): 450 – 460

Abstrak

Carica papaya L. ialah buah tropika yang amat penting di Malaysia bagi memenuhi permintaan tempatan dan eksport. Dalam kajian ini, genom tiga varieti betik (Solo, Eksotika dan Sekaki) telah dijujuk menggunakan platform penjujukan terkini bagi penemuan dan pemilihan polimorfisme nukleotida tunggal (SNP). Himpunan jujukan pendek Solo dan draf genom Sun Up dan Solo digunakan sebagai genom rujukan bagi menjajarkan dua genom betik ini (Eksotika dan Sekaki). Himpunan jujukan pendek Solo dan jujukan contig Sun Up telah menghasilkan 23,318 bilangan scaffolds dengan panjang scaffolds melebihi 200 bilangan nukleotida berpasangan (bp). Penemuan SNP betik menghasilkan 934 SNP putatif dan 7,959 SNP putatif masing-masing dalam Eksotika dan Sekaki. Kajian kami mencadangkan aplikasi teknologi penjujukan terkini (NGS) digabungkan dengan pendekatan bioinformatik bagi membangunkan penanda molekul SNP betik. Penemuan SNP ini membolehkan pengenotipan bertruput tinggi bagi janaplasma betik dan sebagai alat pembiakbakaan berbantukan pemilihan (MAS) dalam program pembiakbakaan betik.